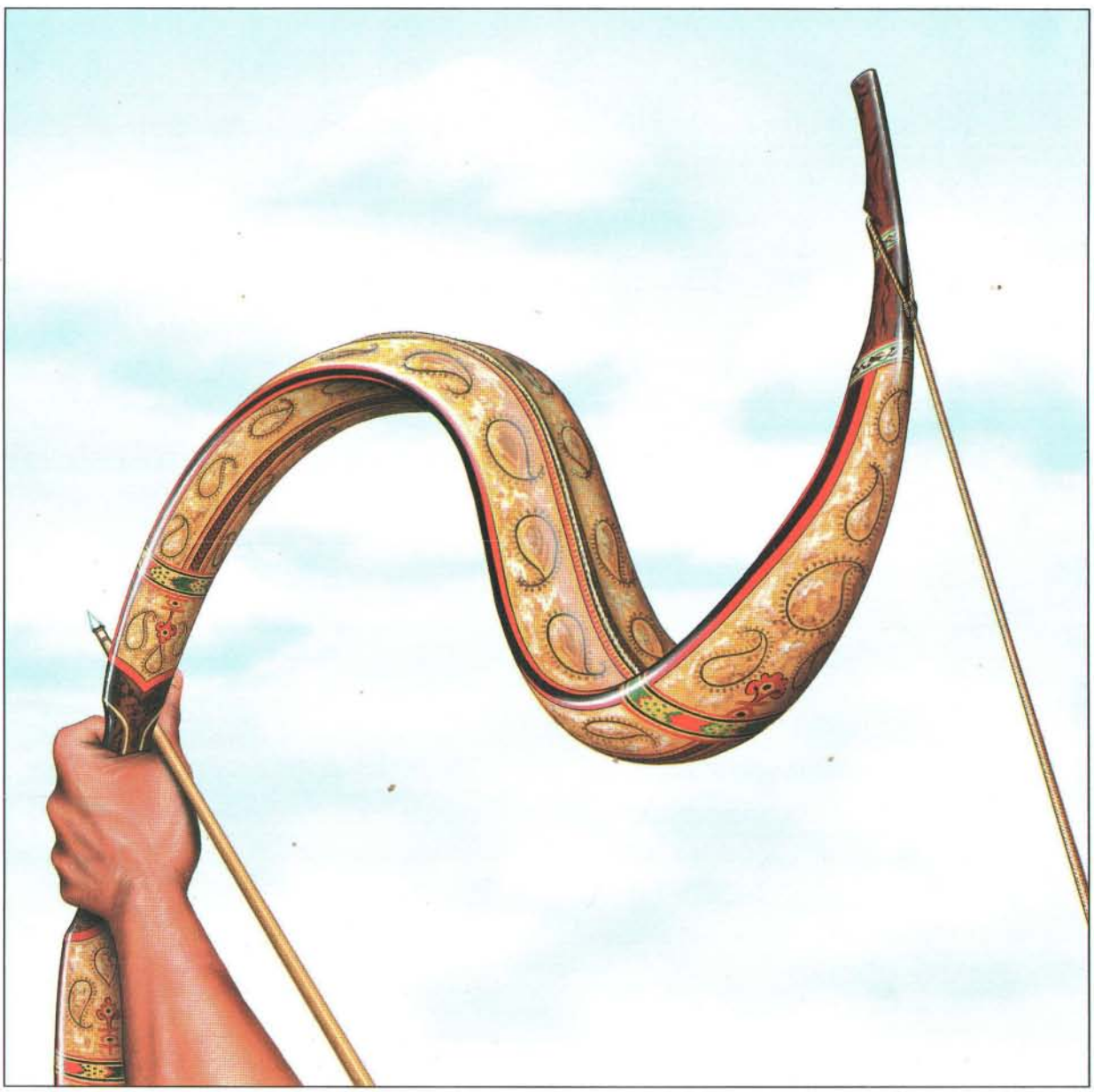
SCIENTIFIC AMERICAN

JUNE 1991
$3.50

Should nuclear power be banned from space?

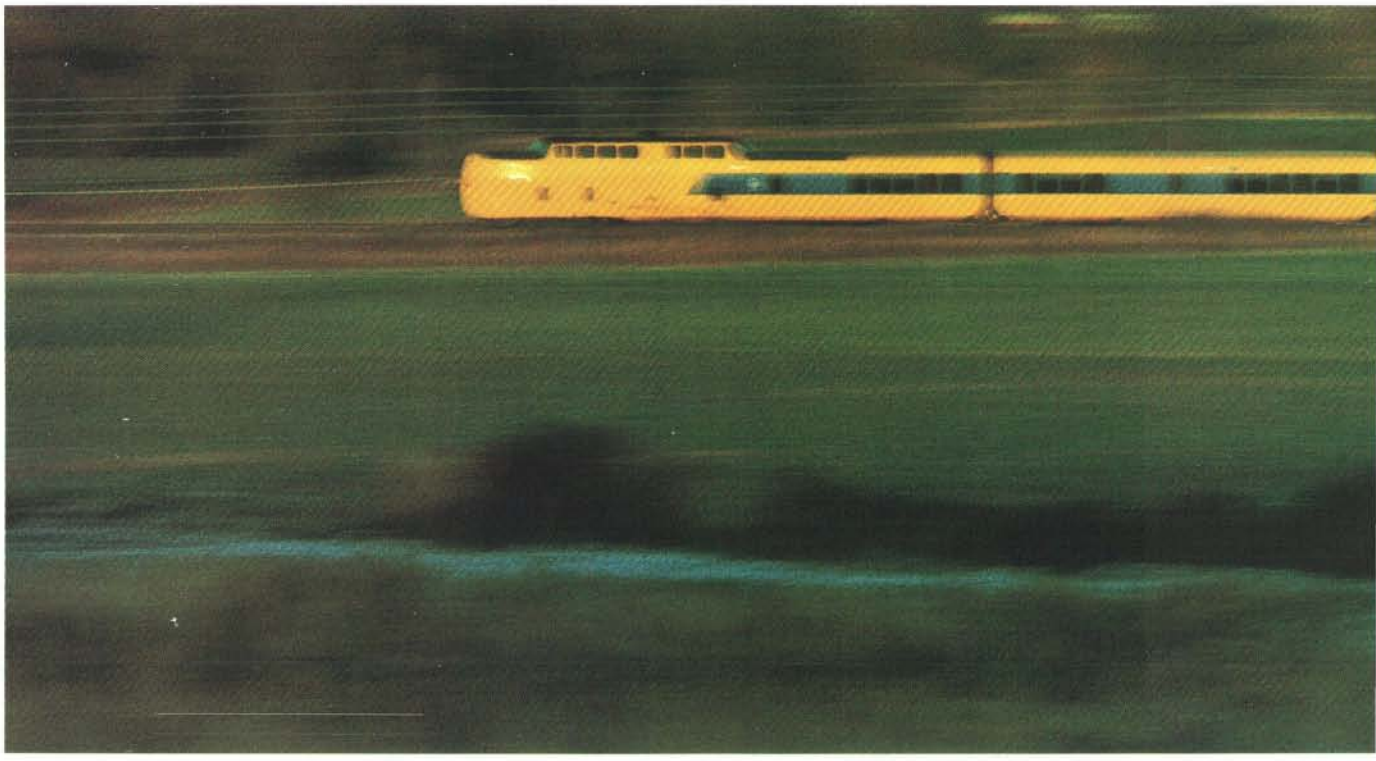
Lasers in the hands of surgeons.

Why quasars can outshine a thousand galaxies.

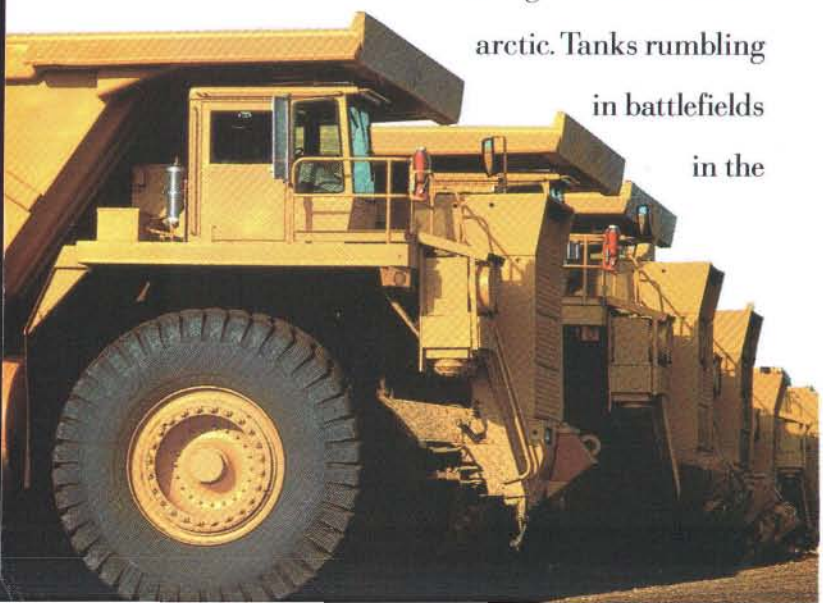


***Early composite bow** could throw an arrow nearly half a mile. Its technology was surprisingly advanced.*

We Can Ke
Whatever Fie

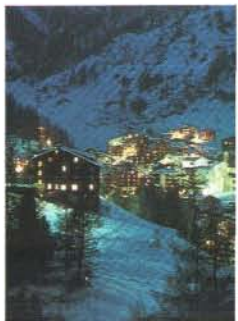


Generators humming in oil fields in the arctic. Tanks rumbling in battlefields in the desert. Trains rolling through fields almost anywhere. Textron Lycoming has the power for them all.

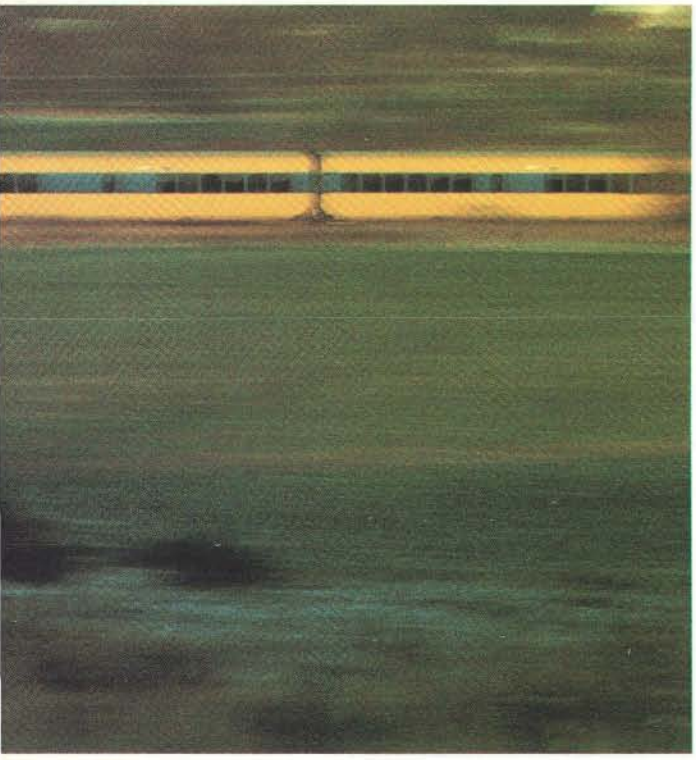


If the power source must be light and compact, clean and quiet, efficient and powerful, the engine should be a versatile Textron Lycoming gas turbine.

They're light enough and compact enough to be mounted on tractor-trailer trucks and used as a



ep You Rolling, ld You're In.



mobile source of emergency power, lighting up an entire village in the snowbound Austrian Alps.

They're clean and quiet enough to run in the august Banco di Roma, providing electricity to power lights and computers during a blackout.

They're efficient enough to be used as a "co-generator" by a public utility in Winnipeg, Manitoba, pumping natural gas to thousands of homes, and using the exhaust heat to power other parts of the plant.

And they're powerful enough to drive 200-ton-payload trucks at a copper mine in Montana; or trains pulling heavy loads over mountains in Europe and Japan.

Textron Lycoming gas turbines are versatile enough to pump liquid gas on the frigid tundra of Canada; natural gas in steamy Pakistan; fresh water in Germany; sea water in Alaska.

If you thought Textron Lycoming only made engines for the M1 Abrams tank or the BAe 146 jetliner, think again. And again. And again. Then call our Advanced Applications

Department at (203) 385-2255. If you've got an idea that needs momentum—on land, at sea, or in the air—call Textron Lycoming, USA.

We're On The Move.

TEXTRON Lycoming

Textron Lycoming/Subsidiary of Textron Inc.

SCIENTIFIC AMERICAN

June 1991 Volume 264 Number 6

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Canadian GST No. R 127387652. Subscription rates: one year $36 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. only 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537.

DEPARTMENTS

THE COVER painting depicts an 11th-century composite bow made in India for hunting and flight shooting. Wrapped in a thin layer of elaborately decorated bark, the bow consists of wood, sinew and horn. The combination of materials creates a weapon more powerful than even the medieval longbow. The engineering of such composite bows made it more practical to shoot an arrow from the side of the bow opposite the hand (see "Early Bow Design and Construction," by Edward McEwen, Robert L. Miller and Christopher A. Bergman, page 50).

THE ILLUSTRATIONS

Cover painting by Hank Iken

Page	Source
19	U.S. Department of Energy
20	Joe Lertola
21	Steven Aftergood (*left*), Joe Lertola (*right*)
22	Johnny Johnson
23	National Aeronautics and Space Administration
25	J.-L. Nieto Observatoire de Toulouse
26	George Retseck
27–28	Edward Bell
29	George Retseck
30	Laurie Grace
31	Edward Bell
33	Vincent A. Fischetti
34	Tomo Narashima
35	Vincent A. Fischetti
36	Johnny Johnson
37	Tomo Narashima
38	Edward Bell
39	Tomo Narashima
40–41	NASA
42–46	Ian Worpole
51	Charles E. Grayson
52–55	Hank Iken
56	Charles E. Grayson
59	Peter Hering, Max Planck Institute for Quantum Optics
60	Andrew Christie; Michael W. Berns (*photograph*)
61	J. Stuart Nelson, Beckman Laser Institute, University of California, Irvine (*top*), Marjorie Mosier, Beckman Laser Institute (*bottom*)
62	Richard Straight, University of Utah
63	Andrew Christie
64	Michael W. Berns (*left*), Aaron Lewis and Daniel Palanker, Hebrew University of Jerusalem (*right*)
67	Royal Astronomical Society Library
68–69	Johnny Johnson
70	Royal Astronomical Society Library
71	W. Eddington
72–73	George V. Kelvin
74–75	Gary A. Glatzmaier, Los Alamos National Laboratory (*left*), Adam M. Dziewonski, Harvard University, and John H. Woodhouse, University of Oxford (*far right*)
76	Jason Küffer
77	Johnny Johnson (*top*), Adam M. Dziewonski (*bottom*)
78	Jeremy Bloxham, Harvard University, and David Gubbins, University of Leeds
79	Jason Küffer
80	Philippe Machetel and Patrice Weber, Groupe de Recherche de Géodésie Spatiale (*left*), Johnny Johnson (*right*)
89–90	Andrew Christie
91	Johnny Johnson

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Alan Hall, *Executive Editor;* Michelle Press, *Managing Editor;* Timothy M. Beardsley; Elizabeth Corcoran; Deborah Erickson; Marguerite Holloway; John Horgan; Philip Morrison, *Book Editor;* Corey S. Powell; John Rennie; Philip E. Ross; Ricki L. Rusting; Russell Ruthen; Gary Stix; Paul Wallich; Philip M. Yam

ART: Joan Starwood, *Art Director;* Edward Bell, *Associate Art Director, Graphics Systems;* Nisa Geller, *Photos;* Johnny Johnson

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Jonathan Goodman; Daniel C. Schlenoff

PRODUCTION: Richard Sasso, *Vice President Production & Distribution; Managers:* Carol Albert, *Prepress;* Tanya DeSilva, *Projects;* Carol Eisler, *Manufacturing & Distribution;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Leo J. Petruzzi, *Manufacturing & Makeup;* William Sherman, *Quality Assurance;* Carl Cherebin

CIRCULATION: Lorraine Leib Terlecki, *Circulation Director;* Cary Zel, *Circulation Manager;* Rosa Davis, *Fulfillment Manager;* Katherine Robold, *Assistant Business Manager*

ADVERTISING: Robert F. Gregory, *Advertising Director.* OFFICES: NEW YORK: William Buchanan; Peter Fisch; Michelle Larsen; Meryle Lowenthal. CHICAGO: 333 N. Michigan Avenue, Chicago, IL 60601; Patrick Bachler, *Advertising Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager;* William F. Moore. WEST COAST: 1650 Veteran Avenue, Suite 101, Los Angeles, CA 90024; Kate Dobson, *Advertising Manager;* Joan Berend, San Francisco. CANADA: Fenn Company, Inc. DALLAS: Griffith Group.

ADVERTISING SERVICES: Laura Salant, *Sales Services Director;* Diane Schube, *Promotion Manager;* Mary Sadlier, *Research Manager;* Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; GWP, Düsseldorf. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie D'Alessandro, *Business Manager*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017
(212) 754-0550

PRESIDENT AND CHIEF EXECUTIVE OFFICER: Claus-Gerhard Firchow

CORPORATE OFFICERS: *Executive Vice President and Chief Financial Officer,* R. Vincent Barger; *Senior Vice President,* Linda Chaput; *Vice Presidents:* Jonathan Piel, John J. Moeling, Jr.

CHAIRMAN OF THE BOARD: Dr. Pierre Gerckens

CHAIRMAN EMERITUS: Gerard Piel

LETTERS TO THE EDITORS

Spies in the Sky

Because several of my books have appeared in the bibliographies of Jeffrey T. Richelson's books, I feel qualified to comment on "The Future of Space Reconnaissance" [SCIENTIFIC AMERICAN, January]. In general, Richelson's effort was accurate and germane, but a number of critical points were either overlooked or ignored. In particular, I take issue with his comment that "high-flying aircraft do not make a good alternative to satellites."

Manned aerial systems have distinct attributes that cannot currently be matched by systems controlled by orbital mechanics. Quick reaction times, for instance, are virtually impossible to achieve with satellites. Points outside conventional orbital parameters can go for days, if not weeks, without being monitored. The predictable arrival of a satellite makes it possible to move classified aircraft under cover to prevent their detection.

Some broad-spectrum electromagnetic sensors are simply too large and cumbersome to be placed in orbit. The manufacturing and operating costs of satellite launch vehicles are often considerably higher than those of aircraft-based systems. And unfortunately, the success rate of satellite launch systems at delivering sensor hardware is significantly lower.

JAY MILLER
President
AeroFax, Inc.
Arlington, Tex.

Richelson responds:

I agree with most of Mr. Miller's statements of fact, but they imply only that aerial reconnaissance is a valuable complement and supplement to space reconnaissance, not an acceptable alternative. The ability of satellites to assure frequent, repetitive coverage of numerous targets over long periods, regardless of their location, cannot be equaled by aerial reconnaissance. When was the last time that a U.S. aircraft put a sensor package over a target in the Soviet Union?

I must also challenge Miller's claim about the relative costs. In the U.S., it is the very existence of an extensive satellite reconnaissance effort that has permitted the operation of a limited-cost aerial effort. If the U.S. had to replace its imaging reconnaissance satellites with enough SR-71 aircraft and related facilities to maintain the level of coverage, I expect that the costs would far exceed those of the satellite program.

Small-Business Blues

The Analytical Economist article "Second-Class Jobs" ["Science and Business," SCIENTIFIC AMERICAN, January] states the problems for the workers in small companies well. The financial burden of attempting to reach parity in salary, health benefits, worker safety and satisfying environmental concerns would tip the balance to unprofitability for many small firms.

Rather than writing off the employees as second class, however, it would be better to seek ways of equalizing their compensation. Government could directly lessen the burden on the employees in several ways, such as adopting an equitable national health insurance. Programs could facilitate and, when necessary, even underwrite compliance with health, safety and environmental regulations. The entrepreneurial risks of small firms pay off broadly for all of us by developing new technologies and identifying business opportunities. And 40 percent of small companies go on to provide stable employment and marketed products!

IVAN G. OTTERNESS
Ledyard, Conn.

Great Men, Small Minds

Timothy Beardsley's profile of Robert C. Gallo ["Science and the Citizen," SCIENTIFIC AMERICAN, January] was illuminating. Predictably, Gallo turned out to be his own worst foe, saying that "the right thing the leader of a group should try to do is intellectually dominate."

The greatest tribute to a leader is the superior quality of his followers. Socrates' finest achievement was Plato, and Plato's was Aristotle. Niels Bohr never appeared to dominate the two dozen first-rate students in the magic circle of his Copenhagen Institute.

By intellectually dominating, rather than stimulating, Gallo surrounds himself with lesser men than himself. The history of science demonstrates that this approach is counterproductive.

JOHN BAKER
Centralia, Wash.

Back in the Harness

As a graduate student in ancient history, I don't often have the chance to get one up on a professor. In "A Roman Factory" [SCIENTIFIC AMERICAN, November 1990], A. Trevor Hodge contends that the inadequacy of the standard harness was such that "if the horse tried to pull a heavy load it merely strangled itself." Recent work, however, has proved that Roman harnessing did not throttle horses.

R. P. BIRD
Wichita, Kan.

Hodge responds:

The view of Roman harnessing I expressed was the standard one of current orthodox scholarship, but it has been disputed by J. Spruytte in *Études Experimentales sur l'Attelage* (Paris, Crépin-Leblond, 1977). That well-documented study has perhaps not had the impact that it should.

Maybe in 2041...

In "50 and 100 Years Ago" [SCIENTIFIC AMERICAN, January], you reprinted an item from 1941 that said a new surgical procedure, the prefrontal leucotomy, "was devised by Dr. Egas Moniz, of Spain." This is not completely correct, since Moniz was Portuguese, not Spanish. His achievements and investigations in medicine brought him the Nobel Prize in 1949.

PEDRO DUARTE FONSECA
Lisbon, Portugal

ERRATUM

On page 89 of "Mathematical Recreations" [SCIENTIFIC AMERICAN, March], certain tangents to Dandelin's spheres are specified incorrectly. The distance between points F and P equals that between P and A. The distance between G and P equals that between P and B.

50 AND 100 YEARS AGO

JUNE 1941

"The fact that only 6,000 light-planes were produced in all of 1940 would seem to indicate that there is not much interest on the part of the public who buys motor cars by the millions. But the reasons for this small demand are the same as they have always been: present-day planes are still relatively difficult to fly; they are dangerous compared to automobiles; and there are a limited number of large landing fields. Ford and United are both aiming to correct these shortcomings by experimenting with planes that can be operated out of small fields—backyards, if you want to stick your neck out that far in prediction. This means that the ultimate plane will be so designed as to be controllable at or very near zero speeds, a possibility with certain helicopter types, as has been amply demonstrated by Igor Sikorsky."

"The flattening of the Earth at the ends of its axis of rotation is caused by the centrifugal force arising from its rotation. There are several quite different ways in which the Earth's shape may be found. We may measure the length of a degree of latitude on the surface; we may measure gravity, which is less at the equator than at the poles, by finding the time of oscillation of a pendulum. Also, the attraction of the Moon on the Earth's equatorial bulge influences the Earth's rotation and causes the precession of the equinoxes. The annual rate of this slow shift of the Earth's axis is very accurately known and can be calculated in terms of ellipticity. Collecting the results from the motions of the Moon and those from gravity on the Earth, Dr. Jeffreys finds that the ellipticity is $1/297.05 \pm 0.38$; that is, that the polar diameter is shorter than the equatorial by this fraction of the latter."

"A skin test which tells in less than an hour whether or not a woman is going to become a mother has been announced by Dr. Frederick H. Falls, Dr. V. C. Freda, and Dr. H. H. Cohen, of the University of Illinois College of Medicine. The test is similar to those made for allergy to hayfever and is said to be 98-percent reliable. Colostrum, a watery liquid secreted in the breasts during pregnancy until milk formation starts, is injected by hypodermic needle into skin of the forearm. If the woman being tested is pregnant, there is no reaction. If she is not pregnant, a reddish area of one or two inches in diameter appears within an hour around the injection point."

SCIENTIFIC AMERICAN

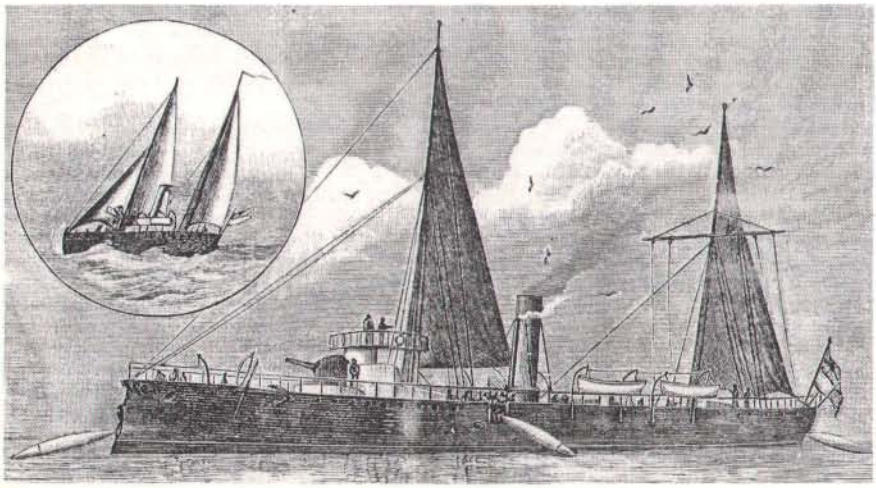
JUNE 1891

"In the annual meeting of the Michigan State Board of Health, Dr. Baker reported that he had worked out the cause of influenza, the prevalence of which has greatly increased during the last three months. He stated that the germs of influenza are generally at all times present but that there must be certain coincident meteorological conditions to irritate the throat and air passages sufficiently to let the germs gain an entrance to the body. These meteorological conditions in this instance were the excessive prevalence of the north and northeast winds, and the excessive amount of ozone during the past three months. Now the causes are known, and the study of the measures for the prevention can begin."

"Mr. Hiram Maxim has for some time past devoted considerable study to the subject of aerial navigation. His practical experiments appear to have crystallized into the form of a machine which might be called a steam kite. The experimental device consists of a thin sheet or kite 4 ft. wide and 13 ft. long, which is propelled by a screw capable of 2,500 revolutions per minute. When properly inclined and pushed forward by the screw at the rate of 30 miles per hour, it will maintain itself in the air; if the forward speed is increased to 35 miles per hour, it begins to ascend; at 90 miles its rising power is quite strong. Mr. Maxim is now at work on a large machine of silk and steel. A petroleum condensing engine will furnish the power."

"Why, asks the *Pall Mall Budget*, is it so difficult and expensive to construct an immense telescope? Some one who is anxious to anticipate events has asked: Why not replace the glass, which is only a medium transmitting light at a different velocity from air, by a properly constructed electric field? It is conceivable that an electric field 50 feet in diameter could be arranged. Just what the nature of this field should be, with our present knowledge, we cannot say, but some day it will be known, and then the secrets of the other planets will be ours."

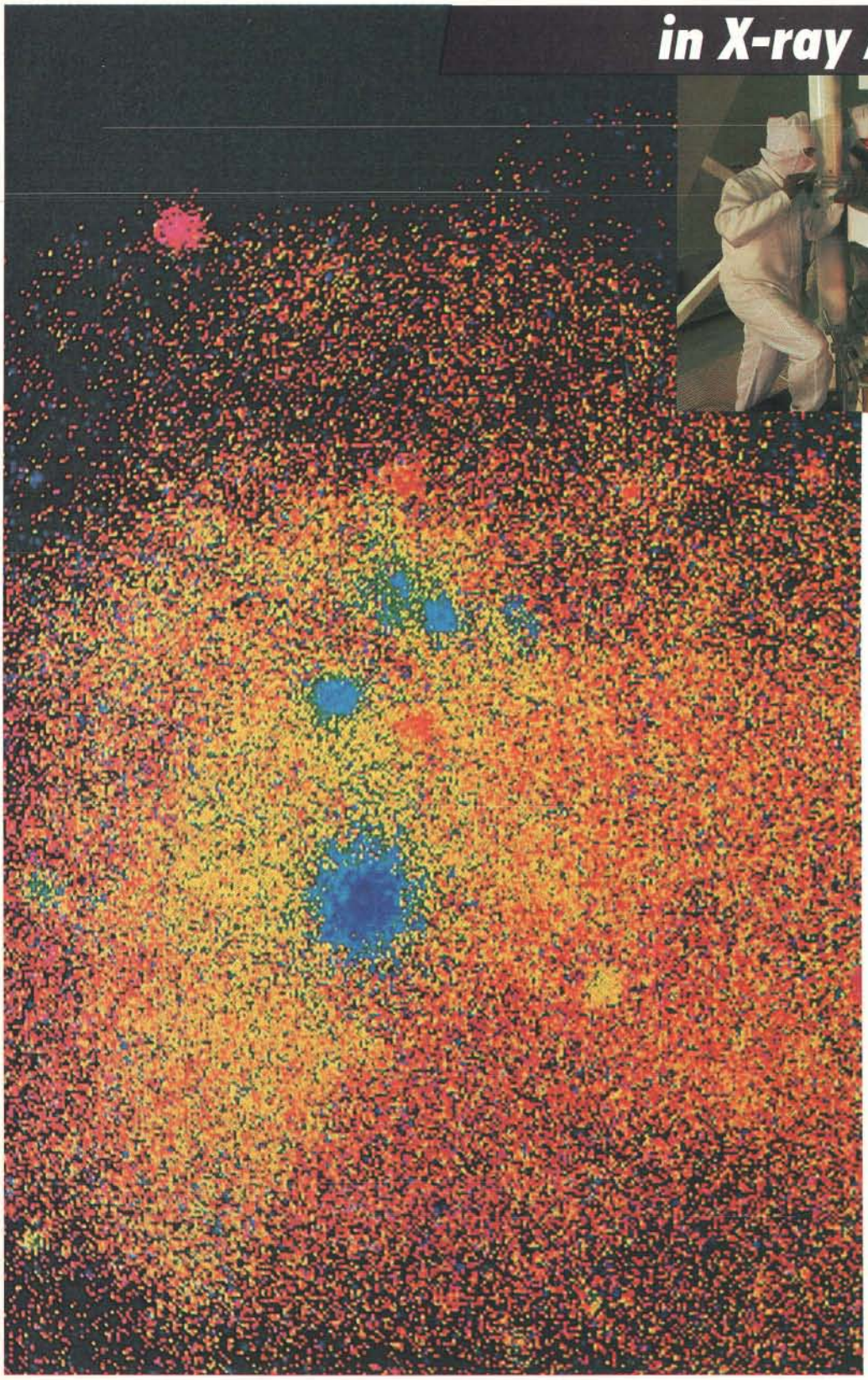
"Among the recent incidents of the civil war in Chile was an engagement in Caldera Bay, a short distance north of Valparaiso, Chile, in which the insurgent vessel, the Blanco Encalada, was sent to the bottom by a torpedo. Our engraving shows a torpedo boat in the act of discharging torpedoes from the low side and the stern."



Launching torpedoes

A New Dimension

in X-ray Astronomy



ROSAT: Hot on the Trail of 100,000 X-ray Sources

The German X-ray satellite ROSAT will enable astronomers and astrophysicists to identify a vast number of new X-ray sources in outer space.

The central role in this major feat of technology will be played by the world's smoothest mirrors: Carl Zeiss produced the mirror system for the satellite's 83-cm Wolter telescope, the largest and most powerful of its kind to date. The extremely low microroughness, unprecedented in a mirror of this type, is 0.25 nm – no more than three times the diameter of a hydrogen atom.

The resolution attained with this mirror system far surpasses anything achieved by X-ray telescopes before.

Large Magellanic Cloud, photographed by ROSAT in X-ray light. The X-ray quadrants are wavelength-coded (blue: short-wave; red: long-wave). By courtesy of Prof. Joachim Trümper, Max-Planck-Institute for Extraterrestrial Physics, Garching.

ZEISS
Germany

Carl Zeiss
D-7082 Oberkochen

Carl Zeiss
Performance and Quality

Science? *Nyet*

Disillusioned Soviets embrace mysticism and the paranormal

A specter never anticipated by Karl Marx is haunting Russia—the specter of occultism. Real specters, too, if you believe the newly freed Soviet press.

Three-eyed extraterrestrials in silvery suits visited southern Russia a year and a half ago, according to Tass, the once staid Soviet news agency. A television faith healer claims to cure not only viewers but also those who drink water or apply cold cream that had been set before the TV screen. A Communist party paper offers young Muscovites astrological advice on the proper timing for sexual activities.

When these reports are picked up in the West, they are presented as equally true by the tabloids, equally false by the elite press and equally amusing by the television anchors. But the rising tide of credulity deeply worries Soviet intellectuals and Western specialists on Soviet affairs. To them the trend signals an ominous decline in the respect accorded to science and rationalism. "A diminishing of the prestige of science, accompanied by a rise in the prestige of occultism, would not bode well for a free society," says Loren Graham, a historian of Soviet science at Harvard University and the Massachusetts Institute of Technology.

To explore the problem, Graham has organized a U.S.-Soviet panel to discuss anti-science trends in both countries. The workshop, to be held at M.I.T. in May, will include scientists and historians from both countries, as well as influential Soviet journalists. They will focus on the Soviet case, however, because they agree that it is the more alarming. "The question is which way is the curve going," Graham says. "Anti-science attitudes are burgeoning in the Soviet Union, but in the U.S. they have been constant for years."

Judging by papers to be read by participants at the M.I.T. meeting, Soviet journalists and scientists themselves are at least partly to blame for the unchecked advances made by the forces of fabulism. "Soviets are not accustomed to a variety of opinions," says Lev Bazhenov, a workshop panelist from the Soviet Institute of Philosophy. "Any kind of nonsense, published in a newspaper, immediately increases its rating significantly."

As superstition gains wider coverage, "it is interesting that at the same time we have a marked decline in the publication of popular science magazines [and the broadcast of] TV programs on science and technology," writes Sergei Kapitza, president of the Soviet Physics Society [see also "Essay," page 96]. Adds Marina Lapina, associate editor of *Science in the U.S.S.R.*: "The blame lies... also with the scientists who always considered the popularization of science of secondary importance."

In extenuation of these failings, some observers point out that *glastnost* came as a shock to a cultural establishment that had never had to worry about its audience. For the first time in 70 years, publications must compete for circulation and even advertising, producing a scramble that often leads editors to place sensationalist stories cheek by jowl with serious accounts. In time, this argument goes, the Russian press will sort itself into high- and low-brow layers, like those found in the West.

Indeed, the garish cover stories on display at the supermarket checkout counters throughout the U.S. are not so different from some Soviet reports. The recent unconfirmed accounts of former Soviet leader Leonid Brezhnev's penchant for astrology hardly surpass Nancy Reagan's famous astrological consultations. Soviet anti-science even gets some of its material from the West. "It is fascinating that in the Soviet Union we are now importing creationism from the U.S. fundamentalists," Kapitza says.

A survey conducted by the Gallup Poll last summer indicated that one in four Americans takes cues from the stars or believes in ghosts. One in six claims to have communicated with the spirits of dead people; one in seven has seen a UFO (unidentified flying object). These figures may seem astoundingly high, but at least they are stable. In fact, the belief in UFOs—the one paranormal belief the pollsters have tracked over time—has actually fallen: only 47 percent affirm the belief, down from 54 percent in 1973.

Data from the Soviet Union, where polling became possible only recently, are still largely anecdotal. No doubt some of the growth in occult opinion is an illusion created by the sudden end of censorship. But experts agree that paranormal beliefs are actually spreading and that regard for science is actually falling.

A decade ago Western Russia-watch-

OPIATE OF THE MASSES? Krishnaites in Moscow. Photo: Sipa/Wallace.

ers interviewed a sample of Soviet emigrés. James R. Millar, a Soviet specialist at George Washington University, who headed that project, says about 80 percent of the emigrés affirmed that science could solve most health problems, for example. Rather lower percentages said science could solve problems of energy, agriculture and pollution.

Today, however, Soviet citizens are less enamored of science than sociologists had expected. Lapina cites polls showing trust in scientific institutions to be lowest among well-educated people—that is, the opinion leaders. It would seem, therefore, that the trend has not bottomed out. "It's a reasonable hypothesis to say there's something of a reaction against the scientism and instrumentalism that was represented by the party," Millar observes.

Many scholars see close parallels with the U.S. experience of the 1960s and early 1970s, when mysticism, anti-scientism and radical environmentalism rose to their current strength on the backs of racial unrest, the Vietnam War and economic dislocation. In the U.S., though, the dislocation was in a time of prosperity, not the dire economic straits now faced by the Soviet populace.

But there are specifically Soviet—or Russian—elements in the story. Not only has the Soviet Union's massive scientific and technological establishment failed to provide a cornucopia, it has littered the country with environmental horrors, above all the 1986 nuclear disaster at Chernobyl. The shock is still sinking in.

Meanwhile the Soviet people have begun to realize that their health care system, long advertised to them and the world as a paragon of equalitarian utility, in fact cannot even supply basic drugs or clean hypodermic needles, except to the well connected. According to Graham, the Soviet Union is the only industrialized country in which the mortality rate has risen during the past 15 years.

The ethnically diverse republics, which long resented being required to support the goals of the central authority, thus have many new reasons to resist paying their share for science, as well as for other cultural activities, such as the ballet. The republic of Georgia, for example, had already stopped picking up the tab for such budgetary items before this spring, when it declared its independence from Moscow.

Paul Josephson, a historian at Sarah Lawrence College, notes that mysticism has deep roots in Russian culture, as evidenced in the 19th-century Slavophile movement, which sought to purge Russia of Western influences. An echo of the Slavophiles could be heard in Aleksandr Solzhenitsyn's speech at Harvard University in 1978, in which he attacked the West for what he called its materialism and soulless rationalism.

Nostalgia for a premodern world has been strengthened by the general backlash against Communism to create a "thirst for other ways of knowing," Graham says. Even Soviet citizens who are not really believers distance themselves from the hated dogma of "scientific socialism"—the official term for Marxism-Leninism—by embracing either orthodox religion or fringe movements, says Lev N. Mitrokhin, deputy director of the Institute of Philosophy.

What is to be done? Lenin, who posed that question in the title of his most famous book, answered it by calling for unquestioning acceptance of Communist discipline. But if democracy is to survive in the Soviet Union, precisely the opposite prescription is needed: the spirit of criticism must be revivified.

Paul Kurtz, a philosopher at the State University of New York at Buffalo who recently visited the Soviet Union, offers a typically Western answer. He plans to establish a Russian version of the *Skeptical Inquirer*, a magazine he founded to investigate reports of paranormal phenomena. "Because of censorship, critical thinking was blocked, not only in the masses but in the elite as well," Kurtz says. "At least in the U.S. we have a significant minority of sophisticated people who consider this all nonsense." —*Philip E. Ross*

Do Screwy Images Solve a Superconductor Mystery?

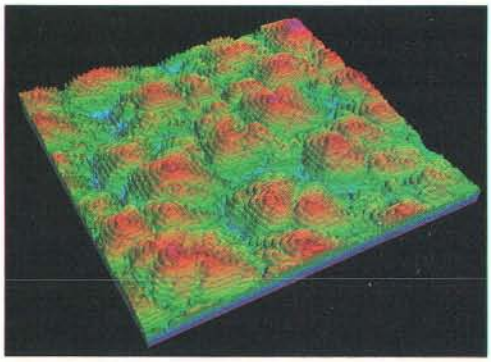
Spiral staircases. Whorls on whorls. Enchanted spiral forest. New microscopic images of high-temperature superconductors have evoked lyrical phrases from *Science* and *Nature*, the scientific journals where the images first appeared. But the pictures are not just pretty. They may help solve a major obstacle to applications of these curious materials.

Just five years ago scientists discovered a class of ceramic compounds that could conduct electricity without resistance at temperatures exceeding 100 kelvins. But efforts to turn the superconductors into highly efficient motors, magnets and other devices have been stymied by their inability to carry much current.

The problem is that current in the ceramics generates magnetic flux lines that "wander" through the material and thereby impede the flow of electrons. This phenomenon is particularly severe in so-called bulk samples of the ceramics, such as wires. Yet thin films of the superconductors, made by condensing vapor on a cold surface, have been strangely exempt from the problem of wandering flux. Researchers suspected that defects in the molecular structure of the thin films "pin down" the flux lines, but the nature of these defects was unknown.

Now scanning tunneling microscope (STM) images made by groups at the IBM Zurich Research Laboratory in Switzerland and the Los Alamos National Laboratory have revealed a defect candidate. The STM images, such as the false-color version from Los Alamos reproduced here, show that thin films consist not of evenly stacked layers but of tiny, spiral-shaped mounds a few hundred angstroms wide at the base.

Both groups suspect that these spirals create "screw dislocations" that can pin down some of the wandering flux lines. Yet the spirals are not distributed densely enough to account for the full current-carrying capacity of the thin films. The Los Alamos workers suggest that flux lines may be pinned down not only by the spirals themselves but also by more subtle defects nestled between them. Resolving this question may help scientists determine how to create defects—and so to improve the current capacity—in other forms of superconducting materials. —*John Horgan*



Homeobox Harvest

A trail of knotted leaves leads to key regulatory genes

When a small bud of tissue in an embryo spontaneously matures into a fully formed arm, it is following orders. Those orders are given by homeobox genes, a family of "master" genes that specify developmental destiny. Homeobox genes are a ubiquitous feature of multicellular animals and fungi—and now it seems of plants as well.

Recently a team of geneticists at the Department of Agriculture's Plant Gene Expression Center and the University of California at Berkeley reported in *Nature* the first discovery of a homeobox gene in plants. Their work further establishes homeobox genes as a universal set of gene regulators that have been telling cells what to become for at least a billion years.

During the 1980s, geneticists and embryologists realized that remarkably complex mutations in fruit flies could result from changes to certain individual genes. These homeotic mutations, as they were called, could remove entire body segments or cause limbs to grow in bizarre places: the *Antennapedia* mutants, for example, had legs where they should have had antennae. The mutated genes appeared to oversee the activity of other genes and, through them, the differentiation of entire sets of cells during development.

Further research revealed that many of the genes responsible for homeotic mutations strongly resembled one another at a certain sequence, which became known as the homeobox. William J. McGinnis, now at Yale University, and other investigators later demonstrated that highly similar homeobox genes could be found throughout the animal kingdom, even in yeast. Yet numerous attempts to find homeoboxes in a wide variety of plants were curiously unsuccessful.

Sarah Hake of Berkeley and her colleagues had not set out to find a plant homeobox, but they were not surprised when their work led them to one. Their studies focused on a mutation in maize called *Knotted*. The leaves of *Knotted* mutants are marred by knots, or swirling protrusions of the lateral veins. The knots "seem to be a group of cells that continue to divide, surrounded by other cells that have stopped," Hake explains. "They're not like tumors. All the cells stay in a nice, neat sheet. It's as though you poked your finger through the material of a sweater to make an outpocketing."

Even the lateral vein cells that have stopped dividing are abnormal: they have characteristics of cells in the sheath, or base of the leaf. Also, in *Knotted* mutants the ligule—a small flap of epidermal tissue normally found between the flat blade and the sheath—grows in the blade itself.

To Hake's group, all the mutant traits seemed to be caused by cells exhibiting normal behavior in inappropriate places or at inappropriate times. As such, they were reminiscent of the homeotic mutations observed in fruit flies. "Ligules in the wrong place aren't as dramatic as *Antennapedia* mutations," she comments, "but in a way, they're analogous."

After several years of work, Hake's group succeeded in identifying the gene responsible for the *Knotted* mutation. They then deduced the amino acid sequence it would make and compared it with those of other known proteins. Sure enough, there were sufficient similarities between the encoded protein and the products of homeobox genes for Hake to classify the *Knotted* gene as one containing a homeobox.

In retrospect, Hake notes, it is clear why previous attempts to find plant homeoboxes failed. The most common search technique involved cross-hybridization, in which DNA strands complementary to one homeobox were used as probes for others. The divergence between the plant and animal DNA sequences is so great, however, that probe molecules based on the animal genes could not adhere to the plant genes reliably. "Now that there's a known plant motif, it is easier to pull out plant homeoboxes," she says. She and her colleagues have already identified more homeoboxes in maize and several other plants, including tomatoes and rice.

Although the plant homeoboxes differ considerably from those found in animals and fungi, the similarities still suggest that all homeoboxes descended from a gene in the organisms' common ancestor, which lived a billion years ago. The role of a homeoboxlike gene in that unicellular ancestor is open to speculation. One possibility, McGinnis notes, is that the gene regulated the transformation of those one-celled creatures into a variety of forms. After multicellularity evolved, that function could have been co-opted to produce different cell types in different body regions.

The discovery of the genes in plants continues to define the range of activities that homeoboxes are known to regulate, adds Matthew P. Scott, a pioneer of homeobox research at Stanford University. "In general, this is a very important type of question: Why is a certain type of regulator molecule involved in a particular process?" he asks. "Is there anything more than an accident of history that determines which of these kinds of proteins is involved?" Or as Shakespeare might have phrased it: Wherefore art thou, homeo? —*John Rennie*

KNOTTED *MAIZE MUTANT* (left) ***differs anatomically from normal plants*** (right). ***An abnormality in a homeobox gene that regulates differentiation causes some leaf cells to express inappropriate characteristics. Photo: Bruce Veit.***

From the Ashes

Will the Gulf War boost the fortunes of Star Wars?

Shortly after the U.S. and its allies began bombing Iraq—and after television viewers had watched U.S. Patriot missiles apparently blasting one Iraqi Scud after another out of the sky—the U.S. launched another offensive on the home front. The objective: to muster support for the Strategic Defense Initiative (SDI), the controversial plan to protect the U.S. from ballistic nuclear missiles.

In his State of the Union Address on January 29, President George Bush proclaimed that the Patriots' performance had shown "we can defend against ballistic missile attacks aimed at innocent civilians." He urged Congress to support a "limited" new SDI plan designed to handle "any future threat to the United States, to our forces overseas and to our friends and allies." In a budget released later, the White House called for increasing next year's funding for SDI by more than 50 percent, from $2.9 billion to $4.6 billion.

Since then, SDI opponents have counterattacked. Some contend that the Patriots' performance has been greatly overestimated and that the missiles may actually have increased casualties in Israel by spreading more debris over a wider area than an unintercepted Scud would have done. More to the point, the critics argue that Bush's new SDI plan is subject to the same flaws as previous ones. "It's the same old wine in a new bottle," says Matthew Bunn of the Arms Control Association.

The limited system proposed by Bush, formally known as Global Protection Against Limited Strikes, or GPALS, is a far cry from the invulnerable "Star Wars" shield envisioned by President Ronald Reagan back in 1983. GPALS could parry attacks of only 200 warheads at most, according to SDI officials. The estimated cost of the system, which the Pentagon hopes to deploy beginning in the mid-1990s, is approximately $45 billion in current dollars, a fraction of the estimates just a few years ago.

GPALS's first line of defense would consist of space-based interceptors called Brilliant Pebbles. Each of these small (one analyst called them "child-sized"), autonomous missiles would supposedly have the ability to detect, home in on and smash enemy missiles. GPALS also calls for deployment of space-based sensors called Brilliant Eyes and of some 1,000 ground-based interceptors—advanced, longer-range versions of the Patriot.

PATRIOTS may not have performed as well in the Persian Gulf War as initially claimed. Photo: D'Essai/Sygma.

SDI officials say such a system could handle three types of attack: an accidental launch by the Soviet Union or some other nuclear power; an unauthorized launch (by a renegade submarine commander, for example, such as the one depicted in the novel and film called *The Hunt for Red October*); or a deliberate launch by another hostile country. Pentagon officials warn that 18 countries now have ballistic missiles, with which they might someday carry out biological and chemical as well as nuclear attacks.

Officially, the cold war is over, and the U.S. no longer fears an all-out nuclear attack by the U.S.S.R., but some SDI personnel seem to feel otherwise. During a recent meeting in New York City, Michael Griffin, deputy director of SDI technology, suggested that the threat of massive retaliation by the U.S. may not be enough to deter the Soviets from launching a first strike. Noting that 20 million Soviets died in World War II, Griffin said "their notion of acceptable losses is different than ours, and I'm worried about that."

Henry F. Cooper, the director of the SDI Office, thinks the U.S. should deploy an SDI system even if it means scrapping the 1972 Anti-Ballistic Missile Treaty, which allows the U.S. and the U.S.S.R. to deploy no more than 100 ground-based ABM missiles at a single site. The Soviets "will scream bloody murder," Cooper acknowledged in an interview with SCIENTIFIC AMERICAN. But he insisted that the Soviets would cooperate if only Congress and the American public would uniformly support the program.

This kind of talk worries Albert Carnesale of Harvard University's J. F. Kennedy School of Government, who helped to negotiate the ABM Treaty. He notes that unilateral deployment of SDI by the U.S. would almost certainly scuttle U.S.-Soviet negotiations aimed at reducing offensive nuclear weapons and could even trigger a new arms race. The Soviets cannot match the U.S. in building a strategic defense, Carnesale says, but they could ensure their ability to retaliate by building more offensive weapons, which are relatively cheap.

Even if this scenario did not occur, GPALS would not be worth the cost, according to Richard L. Garwin, a physicist at IBM and another veteran SDI critic. Garwin points out that both Dick Cheney, the secretary of defense, and Colin Powell, chairman of the Joint Chiefs of Staff, have downplayed the likelihood of unauthorized launches of Soviet missiles and that accidental attacks can be prevented much more easily and cheaply by equipping missiles with radio-controlled disabling devices.

As for the proliferation of ballistic missiles among Third World nations, James Rubin, an arms-control specialist for Senator Joseph R. Biden, Jr., of Delaware, contends that for now the threat to the continental U.S. remains completely hypothetical. "The money could be much more safely and effectively used keeping people from getting this technology," he says.

There is one part of Bush's new SDI plan that even the critics find unobjectionable. It calls for research into upgraded versions of the Patriot: ground- or ship-based defensive missiles that could protect U.S. troops and allies from short- or medium-range ballistic missiles like the Scud. Congress may well provide the $600 million Bush has requested for this effort, says Robert G. Bell of the Senate Armed Services Committee—or even increase it.

Some moderate politicians and independent experts, Bell adds, have also

Of Two Minds about Privacy

The father of computer privacy says he's depressed. Willis H. Ware of the Rand Corporation wrote a report for the U.S. Department of Health, Education and Welfare in 1973 that helped lead to the first rules for personal information in government data bases. Since then, he complains, "I've watched *nothing* happen." Today, he says, "the enemy" is commercial repositories of data rather than government ones, and "privacy has been nickeled and dimed to death" by piecemeal legislation.

Maybe that's just the way Americans want it. According to a survey commissioned by Equifax, one of the three major credit-reporting bureaus, the majority of U.S. citizens may not really mind that people can buy or sell their names, addresses, telephone numbers, salaries, bank balances, purchasing habits or commuting patterns. Speaking at the Conference on Computers, Freedom and Privacy in Burlingame, Calif., this past March, John Baker, the company's senior vice president for consumer and government affairs, contended that most people welcome the fact that credit data bases let them undertake business transactions with perfect strangers.

Equifax's poll, conducted by Louis Harris & Associates, reveals that people have mixed feelings about privacy. Nearly four out of five express general concern about threats to privacy, but the same proportion would be upset if they could not get credit based on their past record of paying bills. Similarly, about two thirds oppose direct-marketing companies being able to buy their personal information, but the same number supports the sale of lists containing the names and addresses of those who might want to receive information about a particular product.

These attitudes translate into actions that are also at cross-purposes. Activists in the U.S. are pushing legislation that would increase people's control over information about them held in other hands. One proposal would force private data bases to send people a copy of their personal files. Meanwhile the computer network Compuserve offers a national directory capable of producing the name and address corresponding to any given telephone number for about 50 cents.

In much of the rest of the industrialized world, people already have rights over data bases containing their personal information. And many practices common in the U.S., such as electronic monitoring of the workplace, are illegal elsewhere. Indeed, Simon Davies of the University of New South Wales calls the U.S. situation "an embarrassment to the privacy movement."

Americans seem to have opted against privacy even in trivial matters: coinless telephones in the U.S., for example, rely on credit cards and a centralized data base, whereas those in Europe typically rely on anonymous magnetic cards that are debited with each call. And electronic highway tollbooths now being tested in several cities take the same approach as the telephone system, recording the identity of all cars passing through them.

In addition, at least two U.S. computer companies are experimenting with "active badges," ID cards containing microchips that can track employees from room to room. Researchers say the information can be used to forward telephone calls automatically or to produce a personal diary of meetings.

The technology already exists, however, to protect privacy instead of encroaching on it. David Chaum of the Centre for Mathematics and Computer Science in Amsterdam, for example, is pushing "digital cash" based on "smart" credit cards and advanced encryption techniques. He says it can guarantee creditworthiness without resorting to central repositories of personal data and buying habits. In fact, Chaum claims his system can replace virtually every other form of personal credentials. The only question is whether anyone will want it if its only advantage is privacy. —*Paul Wallich*

In general, Americans value privacy highly:
79 percent are concerned about threats to personal privacy
79 percent believe that privacy is a fundamental right
71 percent believe that consumers have lost control over personal information
57 percent believe that consumers are asked to provide excessively personal information
30 percent have decided not to apply for a job, credit or insurance to avoid disclosing information
But when it comes to specifics, other priorities come first:
96 percent believe that credit checks for loan applicants are appropriate
94 percent believe that credit checks for credit-card applicants are appropriate
83 percent support preemployment drug testing
78 percent would be upset if they could not get credit based on credit reports
67 percent accept the use of bankruptcy prediction formulas to deny credit. (Of the 25 percent of respondents who reported being denied credit, however, 72 percent believed the denial was unjustified.)
54 percent would allow health insurers to exchange information about previous claims when deciding whether to issue policies
44 percent would be upset if they could not use credit cards to purchase goods and services

SOURCE: Louis Harris & Associates

expressed interest in deploying as many as 300 ABM missiles at several sites in the U.S. Such a system could protect the U.S. from most plausible threats at a relatively small cost and with only minor modifications of the ABM Treaty, according to Michael Krepon of the Henry L. Stimson Center in Washington, D.C. "It won't be amazing science," he says, "but it could be practical." But Krepon says neither he nor—he thinks—a majority of Congress will support the deployment of 1,000 interceptors on the ground and an equal number in space. "That just won't happen," he says.

John E. Pike of the Federation of American Scientists predicts that Congress will agree to fund tactical ABM missiles that improve on the Patriot but will keep a lid on funds for strategic defenses—and the space-based Brilliant Pebbles in particular. He suspects that Bush might even intend this outcome. The GPALS plan, Pike quips, has effectively "deferred the decision on Brilliant Pebbles to the first Quayle administration." Vice President Dan Quayle is said to be the most ardent supporter of SDI in the White House. —*John Horgan*

Double Vision

Binary-optics developers focus on sensors and communications

By etching microscopic grooves in the surface of translucent materials, engineers are creating precise, efficient, inexpensive, paper-thin lenses known as binary optics. More than 50 companies in the U.S., Europe and Japan have begun to capitalize on binary optics. But Wilfrid B. Veldkamp, who helped pioneer the technology at the Massachusetts Institute of Technology's Lincoln Laboratory, believes that, considering the potential of binary optics, researchers have only scratched the surface.

Binary optics can already be found in such systems as infrared cameras, endoscopes and ultraviolet sensors. The lenses have improved the performance of these systems while reducing the number of optical elements, the size of the device and the cost of manufacturing. And because high-quality, binary-optic lenses can be as small as tens of microns in diameter, researchers are discovering that they are well suited as focusing elements for sensor systems, scanning optics for laser printers, read heads for compact disc players and transducers in communications networks.

The first binary-optic lenses had circular, concentric grooves and ridges. Because these features formed a two-level surface, Veldkamp coined the term "binary optics." For a binary-optic lens to function effectively, the depth of the grooves and the spacing between them must be comparable to the wavelength of the light they are intended to modify.

Instead of refracting light like conventional lenses, binary optics diffract light like holograms. The light that strikes the ridges of a binary-optic lens travels through more material than the light that enters the grooves. As a result, the light passing through the ridges is delayed with respect to the remainder, and the light waves diffract, or interfere with one another. The interference effect can be used to focus, filter, polarize, shape, split or combine beams of light. The effect can be modified by altering the pattern of grooves as well as their dimension and shape.

The technology of binary optics is by no means the first to be based on diffraction, but in most circumstances, it is far more efficient than any other. Binary-optic lenses operate best on beams of radiation of a single frequency, for example, laser light. They can be adapted or combined with conventional lenses to handle a broad spectrum of light or to provide a wide field of view.

Most of the principles of binary optics have been known for two decades. What launched the technology in the 1980s were the advances in fabrication that made commercial applications feasible. The first of two manufacturing processes involves a machine that rotates lenses on cushions of air and then cuts them with diamond-tipped tools.

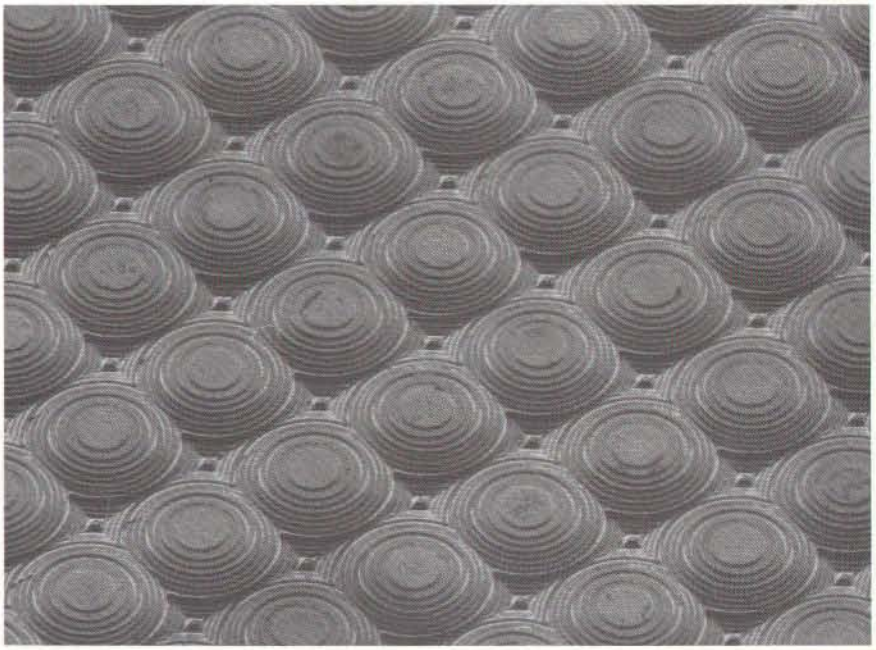
To produce more intricate structures, workers rely on lithographic techniques similar to those used to make integrated circuits: a surface is covered with a patterned mask, and a particle beam etches away the unprotected regions. As many as 20,000 binary-optic lenses can be packed on a surface one square centimeter in area. These microlenses can split or combine light from laser diodes, and in this capacity, they have demonstrated their potential in optical computing.

Microlenses can be combined with microelectronics to make sensors that gather and process data with great facility. H. John Caulfield of the University of Alabama at Huntsville, Veldkamp and many others are developing such sensors for robotic vision and missile guidance. The sensors rely on arrays of microlenses that can focus a small part of a scene onto a detector less than a millimeter away. Indeed, the microlenses concentrate the light so precisely that the detectors can be extremely small, leaving room nearby for processing electronics. The sensors are already capable of crudely detecting the movement of objects or their edges.

Although such high-tech applications foretell a promising future for binary optics, most short-term benefits will come from combinations of conventional and binary optics. Workers at 3M Corporation, for instance, have designed elements to replace the lenses of eyes in patients who suffer from cataracts. Conventional lenses are routinely used in the operation, but they are limited in their ability to allow patients to focus on near or distant objects.

To solve this problem, John Futhey, a senior physicist at 3M, and his colleagues have added a binary-optic element to a conventional lens. As light enters the eye, the element focuses about half of the rays on the surface of the retina and half within the retina. The eye and brain can concentrate on one focal point and ignore the other, providing either near or distant vision. The binary lenses have been implanted by the thousands in 50 countries, and 3M expects that they will be available to surgeons in the U.S. within the year.

Veldkamp wonders whether the optics industry suffers from its own brand of shortsightedness. "I fear that companies in the U.S. are working only on applications that can be developed within the next five years," Veldkamp comments. "To maintain our leadership in this field, we must begin thinking about long-term possibilities that will appeal to a mass consumer market." —*Russell Ruthen*



MICROLENSES advance the effort to integrate optics with electronics. Each lens is 55 microns in diameter, smaller than a grain of sand. Photo: Wilfrid B. Veldkamp.

There's the Rub

Nanotribology reveals the atomic nature of friction

Here's a course of study that has not shown up at many schools yet: nanotribology. It is the study of friction on an atomic scale, a pursuit made possible by the advent of high-resolution microscopy and powerful computers. In their efforts to describe the molecular dynamics of friction, researchers are getting some unexpected—and useful—results.

Uzi Landman and William D. Luedtke, physicists at Georgia Institute of Technology, and Nancy A. Burnham and Richard J. Colton, chemists at the Naval Research Laboratory in Washington, D.C., investigated the molecular interactions that make objects try to stick to one another. They used computer simulations and an atomic-force microscope to see what happens when the tip of a tiny nickel probe approaches a layered substrate of gold. The microscope, which consists of a tip mounted to a cantilever, can measure the forces between two substances separated by less than one angstrom (10 nanometers).

The researchers' systems consisted of 7,000 to 12,000 atoms, so they could "capture the full faithfulness" of the processes unique to the nanoworld, Landman says. "At a certain distance between the tip and surface, about 4.25 angstroms, an instability occurs," Landman explains. Some of the gold atoms "jumped" up to the nickel, adhering to, or "wetting," the bottom of the tip.

Pulling the tip back after it had contacted the surface brought more surprises. A connective "neck" of atoms—a sort of wire on an atomic scale—developed between the two metals. Once the tip is pulled far enough from the substrate, the wire snaps, leaving behind a gold-plated nickel tip and a damaged gold surface.

Landman believes the phenomena can be explained by the different surface energies of nickel and gold. The surface energy refers to the amount of energy it costs for a substance to have an exposed surface. "Gold has a low surface energy compared with nickel, so the system would rather have the nickel surface wetted, or covered, by gold," he says.

The formation of the intermetallic junctions (the necking between the nickel and gold) is the microscopic basis for the macroscopic phenomenon of adhesion and friction. Two surfaces "do not just flatly become glued to each other," Landman observes. "The atomistic mechanisms that bring about the formation of these junctions is fundamental to understanding the resistance to shear—what we call friction."

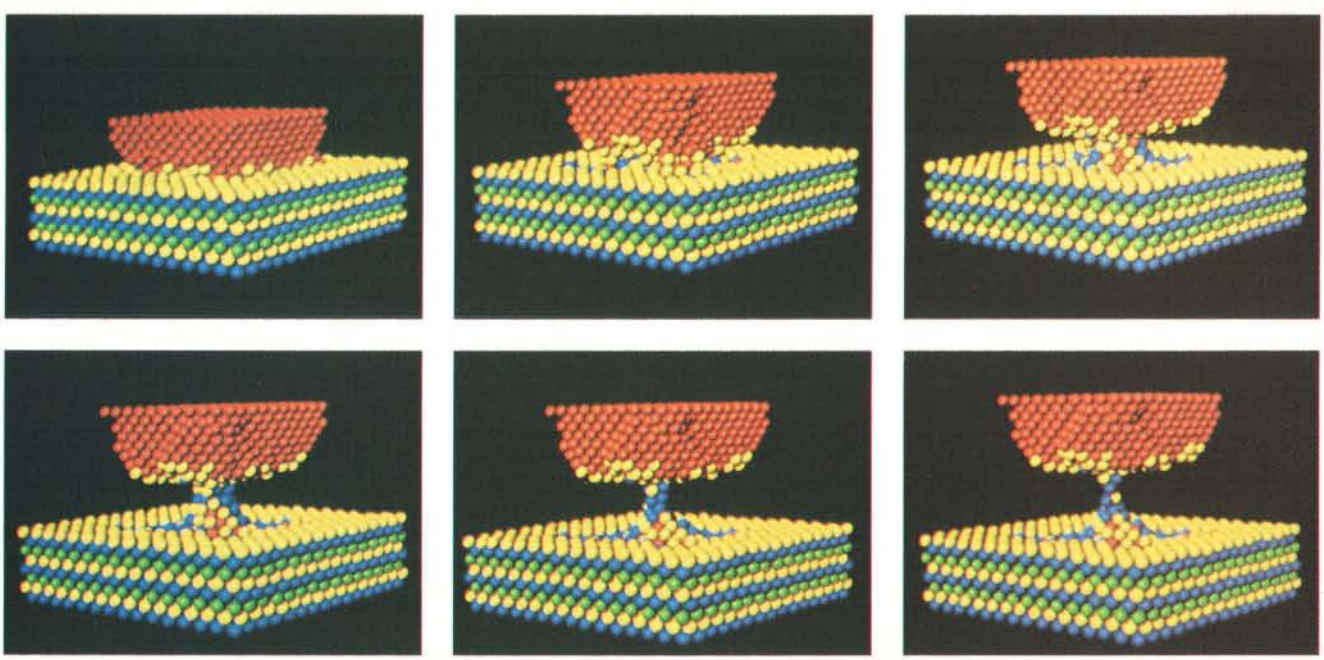
Some nanotribologists believe such studies may soon lead to a "rational design" of lubricants. "The intuition from bulk studies is not very useful," says Jacob N. Israelachvili, a chemical engineer at the University of California at Santa Barbara. Before the nanoscale behavior of materials could be observed, many substances were believed to be good lubricants simply because of their high viscosity.

"But we found it's a totally different reason," Israelachvili says. He and his colleagues studied thin liquid films of hydrocarbons—about one to three atomic layers thick—sandwiched between two surfaces. They found that the lubricating abilities depend on the branching of the chains of molecules. "The molecular geometry is such that the hydrocarbon molecules interlock with neighboring molecules at the surfaces," Israelachvili notes. The lubricant effectively acts as a barrier, preventing the formation of any junctions between the two surfaces.

Understanding friction on a molecular level may also lead to ways of coating substances with thin layers and controlling individual atoms. Now that researchers can manipulate single atoms with a scanning tunneling microscope to spell out letters such as *I, B* and *M,* for example, nanotribology may offer a way to stabilize such configurations, which break down if not kept cold. Scientists may also one day be able to perform "corrective surgery" on microscopic defects in surfaces.

For now, Landman asserts that nanotribologists have finally begun to learn the underlying mechanisms for friction. That means many macroscopic descriptions will "have to be modified to reflect understanding of the small scale."

—Philip Yam



***ATOMIC-LEVEL FRICTION** between a nickel tip* (red) *and a gold surface* (layering of yellow, blue and green) *is described by a computer simulation. The sequence of events begins with the tip indented into the surface* (top left). *As the tip is slowly lifted, an intermetallic junction forms, eventually leading to the "wetting" of the tip by the gold atoms* (bottom right).

Second Guessing

Is more oversight needed to curb scientific cheating?

Since the very nature of science is checking, testing and replicating, it hardly seemed necessary to peer over researchers' shoulders. And, after all, who could be more trustworthy than a seeker of truths? But following recent revelations of apparent fraudulent research at top scientific institutions—and cases involving leading scientists—policymakers are becoming convinced that more must be done to deter misconduct and to ensure an effective response to allegations. "There is a perception that things are out of control," says Erich Bloch of the Council on Competitiveness, who was until last year the director of the National Science Foundation. "And even if that is not true, something more than a business-as-usual approach is needed."

The most publicized recent case began five years ago, when Margot O'Toole, a postdoctoral fellow at the Massachusetts Institute of Technology, questioned data in a paper written by a team that included her supervisor, Thereza Imanishi-Kari, and Nobel laureate David Baltimore, who is now president of the Rockefeller University.

In March 1991, a draft report by the National Institutes of Health Office of Scientific Integrity concluded that Imanishi-Kari had indeed fabricated data both in the original paper, published in *Cell*, and in a subsequent correction letter. Baltimore, whose own contribution to the research has not been challenged, finally retracted the paper. Observers say that if the NIH's findings stand—the draft report was leaked to the press before final review—Imanishi-Kari will probably be barred from receiving federal research funds and may face criminal charges.

But the still unfolding drama could have wider consequences. An oversight subcommittee, chaired by Representative John D. Dingell of Michigan, held a series of hearings on the affair, which Baltimore angrily denounced as a threat to scientific freedom. In April the subcommittee was preparing a potentially damaging report on how M.I.T. and Tufts University, where Imanishi-Kari currently works, investigated O'Toole's challenges. An NIH committee has a similar inquiry under way. As a result, many observers believe the affair could intensify pressure on universities to sharpen their procedures for investigating misconduct.

Science fraud is, of course, not new. Harvard University was shaken by a notorious case in the early 1980s, when John R. Darsee was caught faking research. And the developments in the Imanishi-Kari case follow hard on the heels of another Dingell investigation—into accounting improprieties related to research at Stanford University. Scientific misconduct runs a gamut from such frowned on practices as adding the names of "honorary authors" to research reports to the deliberate fabrication of data.

Still, more cases now seem to be coming to light. Several congressional hearings have examined the frequently heard explanation that science faculty members who consult for high-technology companies are subject to inescapable conflicts of interest. Another view is that intense career pressures on young scientists to "publish or perish" tempt some into cutting corners. Faculty appointments are often made on the basis of research publications, and more are generally seen as better. "Sometimes a scientist falls into sloppiness that evolves into outright misrepresentation," says Suzanne W. Hadley, who was in charge of the NIH's investigation of Imanishi-Kari.

In an attempt to keep its own house in order, Harvard introduced guidelines on research conduct in 1988 to combat the publish-or-perish syndrome. The guidelines suggest, for instance, that only five publications should be assessed for junior appointments. They also stipulate that in collaborations, principal authors review all data. Other universities have adopted similar policies.

A committee of the National Academy of Sciences, chaired by Edward E. David, Jr., a science adviser to President Richard M. Nixon, is considering whether such guidelines should be promulgated for all scientists. David voices popular sentiment in scientific circles when he says, "You're better off if contentious cases are handled close to the people involved. I would look to local mechanisms first."

But even supporters of that view believe that self-policing by the universities is far from easy. Eleanor Shore, dean for faculty affairs at Harvard Medical School, points out that a determined scientific cheat is unlikely to be deterred by rules. The committees convened by M.I.T. and Tufts to investigate, for example, the allegations against Imanishi-Kari found no evidence of wrongdoing. David concedes that many universities lack the expertise to conduct complex investigations fairly—and promptly. Hadley has similar doubts. "It's a real paradox that in some of the most difficult, challenging and sensitive situations, we put the responsibility back on the institutions," she says.

Hadley notes, for example, that when university officials call to say they are initiating an investigation of misconduct, she usually has to impress on them the need to secure immediately all primary data. A bungled investigation can destroy the career of a wrongly accused researcher or of a whistleblower whose complaint is dismissed. (O'Toole was fired after making her allegations and was unable to find work in science for four years.)

Moreover, despite the recent spate of publicized cases, ill-founded charges are more likely than out-and-out fraud. The NIH Office of Scientific Integrity, which investigates NIH-funded research when a local inquiry has found wrongdoing or when a whistleblower is dissatisfied, has concluded 110 investigations since it was created in 1989. It found misconduct in only 16 of them.

The NIH's procedures, however, also have their critics. Barbara F. Mishkin, an attorney who represents several accused scientists, argues that the NIH denies due process to those under suspicion. Mishkin points out that leaked draft reports from the Office of Scientific Integrity can—as in the Imanishi-Kari case—damage reputations before the findings can be challenged.

Bloch and a growing number of other science administrators think what is needed is "a mechanism so that all parties, including the institutions, can obtain help" in resolving allegations. Bloch emphasizes that he would like to see scientists themselves enact such a mechanism, to avoid chilling free inquiry. But Peter Stockton, a staff member of Dingell's subcommittee, believes the root of the problem is unwillingness, rather than inability, to investigate. Both M.I.T. and Tufts had procedures for dealing with allegations of misconduct, he suggests, but they failed to uncover the facts.

One option being studied is that Congress may punish universities that are lax, perhaps by barring them from receiving federal funds. "That is the key," Stockton says. "The procedures are worthless unless you have the will to do it."

It is certain that more instances of misconduct will come to light—especially with the payoff in publicity for investigators. The NIH has several probes under way, and the NIH fraud busters who pressed the Imanishi-Kari investigation are now using computers to search for plagiarism. They say they have already netted a big case. —*Tim Beardsley*

Questioning the "It from Bit"

It's hard keeping up with John Archibald Wheeler. When we leave his third-floor office at Princeton University to get some lunch, he spurns the elevator—"Elevators are hazardous to your health," he declares—and charges down the stairs. He hooks an arm inside the banister and pivots at each landing, letting centrifugal force whirl him around the hairpin and down the next flight. "We have contests to see who can take the stairs fastest," he says over a shoulder.

Outside, Wheeler marches rather than walks, swinging his fists smartly in rhythm with his stride. He pauses only when he reaches a door. Invariably, he gets there first and yanks it open for me. After passing through, I wait a moment in reflexive deference—after all, the man will be 80 years old in July—but a moment later he's past me, barreling toward the next doorway.

The metaphor seems so obvious I almost suspect it is intentional. Wheeler, a professor emeritus of physics at Princeton and the University of Texas at Austin, where he holds a joint appointment and spends a few weeks each year, has made a career of racing ahead of other scientists and throwing open doors for them. He has helped gain acceptance—or at least attention—for some of the most outlandish ideas of modern physics, from black holes to multiple-universe theories. "He has this great ability to see what is important before anyone else and persuade others that this is so," says David Deutsch, a physicist at the University of Oxford.

Wheeler is also renowned for his coinages, analogies and aphorisms, both self-made and co-opted. Among the one-liners he bestows on me are, "If I can't picture it, I can't understand it" (Einstein); "Unitarianism [Wheeler's official religion] is a feather bed to catch falling Christians" (Darwin); "Never run after a bus or woman or cosmological theory, because there'll always be another one in a few minutes" (a professor of French history at Yale); and "If you haven't found something strange during the day, it hasn't been much of a day" (Wheeler).

Lately Wheeler has been drawing his colleagues' attention to something strange indeed. It is a worldview uniting information theory, which seeks to maximize the efficiency of data communications and processing, with quantum mechanics. As usual, Wheeler has packaged the concept in a catchy phrase: "it from bit." And as usual, he delights in being ahead of—or at least apart from—the pack. "I hope you don't think I'm too much like Daniel Boone," he says slyly. "Any time someone moved to within a mile of him, he moved on."

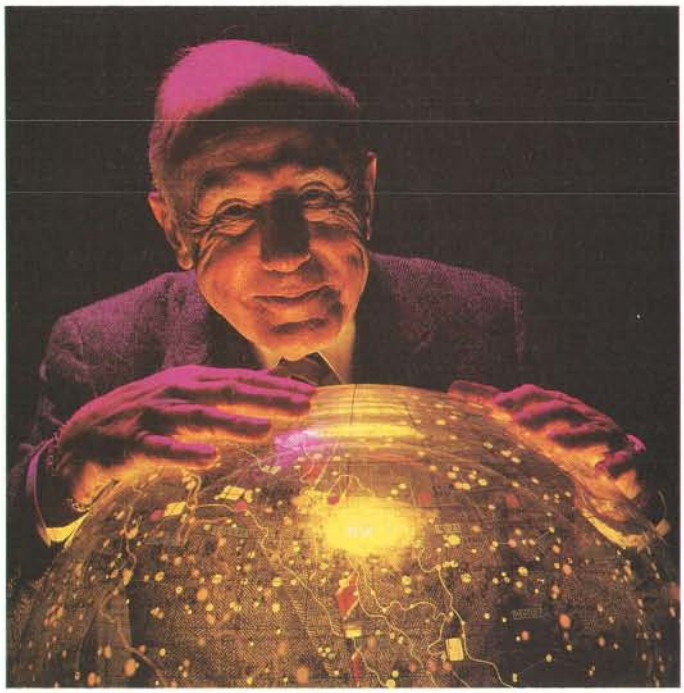
Wheeler might have been dismissed as fun but flaky long ago if he did not have such unassailable credentials. The son of two librarians "who were interested in ideas, interested in the world, interested in adventures" (and who obviously endowed him with an omnivorous appetite for reading), he entered Johns Hopkins University at the age of 16 and emerged with a Ph.D. in physics six years later.

He subsequently journeyed to Copenhagen to study with Niels Bohr, the great Danish physicist, "because he sees further ahead than any man alive," Wheeler wrote on his application for the fellowship. In 1939 Bohr and Wheeler published the first paper successfully explaining nuclear fission in terms of quantum physics. Wheeler's expertise in nuclear physics led to his involvement in the construction of the atomic bomb during World War II and, during the cold war's early years, the hydrogen bomb.

I had heard that beneath Wheeler's puckish demeanor lay a core of steel. That is apparent when I ask if he has any second thoughts about helping to create nuclear weapons. His eyes narrowing, he acknowledges that "a lot of my friends have gone around giving what I call 'scare-the-dope speeches' " deploring such weapons. But he has no regrets. Nuclear weapons, he insists, saved lives by ending World War II quickly and by deterring Soviet aggression thereafter.

When his involvement in the H-bomb project ended, Wheeler immersed himself in studying relativity and gravity—which he calls his "lifelong love"—at Princeton. In 1966 he proposed that a brilliant cloud of gas known as the Crab nebula was illuminated from within by a whirling sphere of solid neutrons created by the implosion of a star. Astronomers later detected such a spinning neutron star, or pulsar, both in the Crab nebula and elsewhere in the Milky Way.

Wheeler also speculated that matter could collapse even beyond the solid-neutron state, becoming so dense that nothing—not even light—could escape its gravitational clutches. Such an object was first proposed by J. Robert Oppenheimer and Hartland S. Snyder in 1939, but it had been dismissed as a theoretical curiosity and not something that might actually exist.



Wheeler's cosmos is participatory. Photo: J. Pinderhughes.

Wheeler recalls discussing such "completely collapsed gravitational objects" at a conference in 1967, when someone in the audience casually dropped the phrase "black hole." Wheeler immediately adopted the phrase for its brevity and "advertising value," and it caught on. Largely because of Wheeler's proselytizing, black holes now play a crucial role in astrophysics and cosmology.

In the 1950s Wheeler grew increasingly intrigued by the philosophical implications of quantum physics. According to quantum theory, a particle such as an electron occupies numerous positions in space until we observe it, when it abruptly "collapses" into a single position. Wheeler was one of the first prominent physicists seriously to propose that reality might not be a wholly physical phenomenon. In some sense, Wheeler suggested, reality grows out of the act of observation and thus consciousness itself: it is "participatory."

These ruminations helped to inspire two of the odder notions of modern physics. In 1957 Hugh Everett III of Princeton, in a Ph.D. thesis supervised by Wheeler, proposed the many-worlds theory: although we can observe a particle in only a single position, it actually occupies all the positions allowed it by quantum theory—in different universes. Four years later another Princeton physicist, Robert H. Dicke, introduced the anthropic principle: it asserts that the universe is the way it is because if it were not, we would not be here to observe it. Although many physicists recoiled from such ideas as untestable and therefore unscientific, Wheeler urged that they be taken seriously.

At the same time, Wheeler began to draw his colleagues' attention to some intriguing analogies between physics and information theory, which was first proposed by Claude E. Shannon of Bell Laboratories in 1948. Just as physics builds on an elementary, indivisible entity that depends on the act of observation—namely, the quantum—so does information theory. Its "quantum" is the binary unit, or bit, which is a message representing one of two choices: heads or tails, yes or no, zero or one.

In addition, information theory provided a new way of viewing entropy, one of the most important, and confusing, concepts in physics. Entropy is defined as the disorder, or randomness, or "shuffledness," as one physicist has put it, of a system. Shannon had proposed that the information in a given system—the sum total of all its possible messages—is a function of its entropy; as one increases, so does the other. Wheeler pointed out that entropy, like a quantum event, is thus tied to the state of mind of the observer. The potential information of a system is proportional to one's ignorance, and so, therefore, is the entropy of the system.

Wheeler was not the only scientist to recognize these links, "but he was probably the first to recognize the potential implications for fundamental physics," says physicist Wojciech H. Zurek of Los Alamos National Laboratory. In the early 1970s Wheeler's speculation bore some tangible fruit when yet another of his graduate students, Jacob Bekenstein, described a black hole in terms of information theory. The surface area of a black hole's "event horizon," Bekenstein showed, is equal to its thermodynamic entropy, which in turn is equal to the information that the black hole has consumed.

Spurred by this and other findings, an ever larger group of researchers—including computer scientists, astronomers, mathematicians and biologists as well as physicists—has passed through the doors flung open by Wheeler. In the spring of 1989 a number of them gathered at the Santa Fe Institute in New Mexico to update one another on their progress. The proceedings of the meeting have just been published by Addison-Wesley as *Complexity, Entropy and the Physics of Information*.

The lead chapter of the book is based on Wheeler's address to the meeting, and it is vintage Wheeler. Over the course of 16 pages, he cites 175 sources, including the Greek poet Parmenides, Shakespeare, Leibniz, Einstein and graffiti in the men's room of the Pecan Street Cafe in Austin, Tex., which states: "Time is nature's way to keep everything from happening all at once." Wheeler also spends some time establishing what reality is not: it is not a "giant machine, ruled by any pre-established continuum physical law"; at its most fundamental level, it even lacks dimension, such as space or time.

Wheeler thinks the whole show may be explained by the "surprise" version of 20 questions.

What is reality, then? Wheeler answers his own question with the koan-like phrase "it from bit." Wheeler explains the phrase as follows: "Every it—every particle, every field of force, even the spacetime continuum itself—derives its function, its meaning, its very existence entirely—even if in some contexts indirectly—from the apparatus-elicited answers to yes-or-no questions, binary choices, bits."

Elaborating on this idea, Wheeler evokes what he calls the "surprise" version of the old game of 20 questions. In the normal version of the game, person A thinks of an object—animal, vegetable or mineral—and person B tries to guess it with a series of yes-or-no questions. In surprise 20 questions, A only decides what the object is *after* B asks the first question. A can then keep choosing a new object, as long as it is compatible with his previous answers. In the same way, Wheeler suggests, reality is defined by the questions we put to it.

How do other scientists react to such propositions? Zurek, who organized the Santa Fe meeting and edited the proceedings, calls Wheeler's style "prophetic, leading the way rather than relating what's already been done."

Wheeler acknowledges that the ideas of the entire field are still raw, not yet ready for rigorous testing. He and his fellow explorers are still "trying to get the lay of the land" and learning how to converse in the language of information theory. Wheeler says the effort may lead to a powerful new vision of "the whole show" or to a dead end. "I like that phrase of Bohr's: 'You must be prepared for a surprise, a very great surprise.'"

Another favorite Wheelerism is "one can only learn by teaching." Wheeler has been the supervisor for some 50 Ph.D.'s in physics during his career, an "enormous number," according to Jeremy Bernstein, a physicist and science writer. Wheeler's most famous student was the late Richard P. Feynman, who received a Nobel Prize in 1965 for his work in quantum electrodynamics. Technically, Wheeler can teach no longer. "If you know of a school that lets its professors teach after they reach 70," he says, "let me know."

But of course, Wheeler can neither stop teaching nor stop learning. During my visit, we run into a young physicist who briefs us on his new cosmological theory, which posits that the universe is riddled with knotlike spatial "defects." "I can't believe space is that crummy," Wheeler declares. Noting the physicist's somewhat crestfallen expression, Wheeler touches his arm and says: "To hate is to study, to study is to understand, to understand is to appreciate, to appreciate is to love. So maybe I'll end up loving your theory." The smile returns to the young man's face, and Wheeler marches off.

—*John Horgan*

SCIENTIFIC AMERICAN

Nuclear Power in Space

The best course for space-borne reactors? Ban them from Earth orbit and use them in deep space, the authors say

by Steven Aftergood, David W. Hafemeister, Oleg F. Prilutsky, Joel R. Primack and Stanislav N. Rodionov

Space nuclear power is a double-edged sword. Although it has played a constructive role in the exploration of space and could continue to do so, it has been burdened by an extensive history of accidents and failures, both Soviet and American. Numerous nuclear-powered spacecraft have released radioactive materials. Spent reactors now in Earth orbit exacerbate the threat posed by orbital debris. And radiation from orbiting reactors has interfered with the operation of other satellites.

In addition, nuclear power in space has in general been a source of international tension because of its role in Soviet and American military space programs. As a result, organizations of both Soviet and American scientists (of which we are members) have proposed banning the use of nuclear power in Earth orbit. Such a ban would reduce the risks associated with nuclear power in space, while permitting its use in those deep-space missions for which nuclear power is essential.

The first nuclear-powered spacecraft was *Transit 4A*, a navigational satellite launched by the U.S. in 1961. *Transit* used a radioisotope thermoelectric generator (RTG), which converted heat from decaying plutonium 238 to electricity. The first nuclear accident in space came less than three years later: *Transit 5BN-3*, the fifth RTG-powered craft to be launched, failed to achieve orbit in April 1964. Its power source disintegrated in the atmosphere (as early RTGs were designed to do) at an altitude of about 50 kilometers. Release of its 17,000 curies of plutonium 238 fuel tripled the worldwide inventory of this isotope and increased the total world environmental burden of all plutonium isotopes—mostly from atmospheric testing of nuclear weapons—by about 4 percent. (Current RTGs contain an order of magnitude more radioactive material.)

In 1965 the U.S. launched its first and only space nuclear reactor, the prototype SNAP-10A. (Reactors generate heat by means of a controlled fission chain reaction rather than simple radioactive decay.) The SNAP-10A operated for 43 days; it is still in orbit. Later that same year the U.S.S.R. sent its first RTG-powered satellites into space. It also launched radioisotope-powered *Lunokhod* lunar modules in 1969 and 1973.

After 1970, however, the Soviet program largely revolved around the *Radar Ocean Reconnaissance Satellite* (*RORSAT*), used to monitor U.S. naval forces. The small reactors on board these craft produce about two kilowatts of electricity. Although a *RORSAT*'s power requirements could be met by solar panels, the craft uses a reactor because solar panels would cause too much drag at the *RORSAT*'s typical altitude of around 250 kilometers. The limited range of the craft's radar necessitates the low orbit.

After a lifetime of several months, a *RORSAT*'s reactor is ordinarily boosted to a "disposal orbit" at approximately 950 kilometers, while the body of the spacecraft reenters the atmosphere. The disposal orbit of the reactor guar-

STEVEN AFTERGOOD, DAVID W. HAFEMEISTER, OLEG F. PRILUTSKY, JOEL R. PRIMACK and STANISLAV N. RODIONOV represent working groups of the Federation of American Scientists and the Committee of Soviet Scientists for Global Security, which have jointly proposed a ban on the use of nuclear power in Earth orbit. Aftergood is a senior research analyst at the Federation of American Scientists in Washington, D.C. He received his B.Sc. in engineering from the University of California, Los Angeles, in 1977. Hafemeister is professor of physics at California Polytechnic State University and a professional staff member of the Senate Foreign Relations Committee. He has worked on arms-control treaties both in the State Department and the Senate. He received a doctorate in physics from the University of Illinois at Urbana-Champaign in 1964. Prilutsky is a space physicist and department head at the Space Research Institute of the Soviet Academy of Sciences. He received his Ph.D. in physics from the Moscow Physical Engineering Institute in 1973. Primack is professor of physics at the University of California, Santa Cruz. A specialist in theoretical particle physics and cosmology, he received his Ph.D. in physics from Stanford University in 1970. Rodionov is a high-energy physicist and laboratory head at the Space Research Institute of the Soviet Academy of Sciences. He received his Ph.D. in physics from the Nuclear Physics Institute in Novosibirsk in 1958. The authors gratefully acknowledge the contributions of Daniel O. Hirsch, Roald Z. Sagdeev and Frank von Hippel.

CLEANING UP AFTER *RORSAT*: a Soviet surveillance satellite (*Cosmos 954*) reentered the earth's atmosphere over the Northwest Territories in 1978, littering radioactive debris over thousands of square miles. In the photographs above, workers gather both large and small fragments of the satellite and its reactor. Decontamination cost the Canadian government approximately $10 million. Proposed U.S. nuclear-powered spacecraft would produce hundreds of times as much radioactivity.

The Next Step in Space Nuclear Power?

There are two fundamental sources of nuclear power for applications in space: reactors and radioisotope power supplies. Whereas a reactor produces heat through the controlled fission of uranium fuel, a radioisotope thermoelectric generator, or RTG, derives heat simply from the decay of a highly radioactive material. In both cases, the heat is converted to electric power. The RTG is best suited for power requirements of less than a few kilowatts, the reactor for higher power levels.

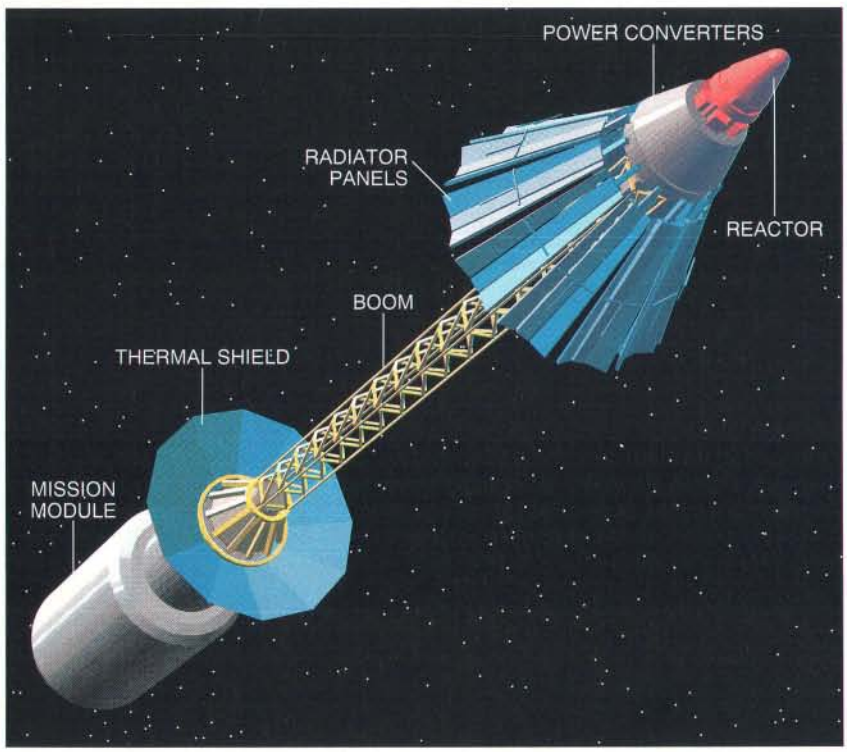
Although the U.S. has launched only one nuclear reactor into orbit, an ambitious reactor development project has been under way for most of the past decade [see *illustration below*]. As currently planned, the SP-100 reactor would generate approximately 100 kilowatts of electricity from 2.5 megawatts of thermal power—far more power than any reactor flown to date. It would contain about 190 kilograms of uranium nitride fuel enriched to 96 percent in the fissionable isotope uranium 235.

The entire reactor is intended to weigh approximately 3,000 kilograms, a mass-to-power ratio of 30 kilograms per kilowatt. Except for a small "shadow shield," which helps to protect the payload from the intense radiation emitted during operation, the SP-100 is designed to be unshielded.

The reactor would be cooled by liquid lithium metal, which would flow through pipes to thermoelectric cells—circuits containing junctions between dissimilar metals that can transform a temperature difference into a voltage difference. These cells would convert about 4 percent of the reactor-generated heat into electricity. The considerable waste heat would be ejected through a set of radiator panels with a surface area of around 100 square meters.

Nearly every component of the SP-100 design extrapolates beyond existing technological experience, making it uncertain whether the program will be able to achieve its goals. Moreover, the reactor has been designed in the absence of a proposed mission, and so any specific application is likely to require substantial revision.

Flight-testing of the SP-100 will not be possible before the turn of the century; cost estimates for the test alone exceed $1 billion, a discouragingly large sum. Furthermore, as the program moves toward its second decade, a specific mission for the SP-100 has still not been defined.



antees that it will not reenter the atmosphere for several hundred years, by which time most of its radioactivity will have decayed.

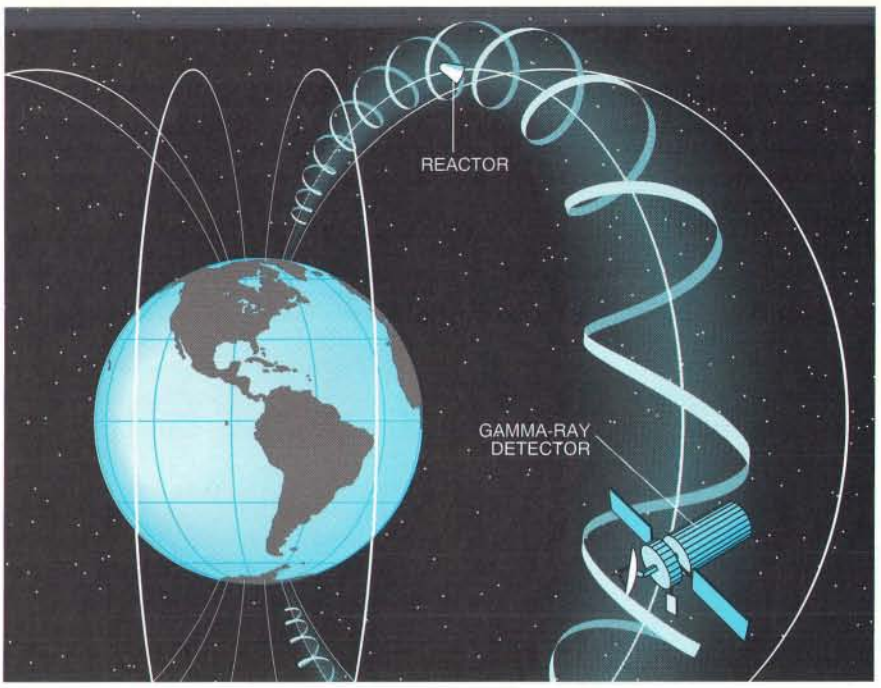
Between 1970 and 1988 the Soviets deployed 31 *RORSAT*s. The jettisoned reactors from 29 of these systems are still in orbit, along with hundreds of kilograms of radioactive fuel. Two have reentered accidentally. *Cosmos 954* fell to the earth on January 24, 1978, spreading radioactive debris over thousands of square miles of northwestern Canada. U.S. President Jimmy Carter immediately proposed a ban on orbiting nuclear-powered satellites, but the U.S.S.R. did not respond.

After the *Cosmos 954* incident, the Soviet Union introduced a backup fuel-core ejection system, so that in the event of accidental reentry, the fuel would disintegrate in the atmosphere. This measure increases the total population that may eventually be exposed to radiation, but it minimizes the exposure received by any single person. In 1983 the fuel core on board *Cosmos 1402* at least partially disintegrated in the upper atmosphere over the South Atlantic Ocean.

Another *RORSAT, Cosmos 1900,* ceased to respond to radio commands in April 1988. Days before its anticipated reentry, however, an automated backup booster system was successfully activated, and the reactor was lifted to a high orbit. The Soviet Union has not launched any additional *RORSAT*s (or other space nuclear power systems) since then.

In 1987 the Soviets flight-tested the Topaz, a next-generation reactor. One flew on board *Cosmos 1818* and operated for six months; the other, on board *Cosmos 1867,* operated for a year. Each generated about 10 kilowatts of electricity. The Soviets have since offered Topaz reactors for sale to Western nations, and the U.S. government has purchased an unfueled version for ground testing and development.

It is possible to detect a learning curve in the history of nuclear power in space. The identical type of accident has never happened twice. Designers have learned such obvious lessons as encapsulating RTG power sources so that they can withstand most launch failures and accidental reentries. Reactors, too, have been redesigned to improve safety. Nevertheless, the power levels, operating lifetimes and radioactive payloads of current and projected space nuclear power systems have also increased significantly. Future accidents could therefore have unacceptable consequences.



UNSHIELDED ORBITING REACTOR emits a cloud of electrons and positrons that spiral around the earth's magnetic field lines and create a temporary radiation belt. A satellite passing through the belt is subject to bursts of gamma rays as the positrons annihilate electrons in its outer skin. Such bursts have disrupted the operation of astronomical satellites.

TOPAZ REACTOR, flight-tested by the Soviet Union in 1987, has been purchased by the U.S. for ground testing. The Topaz is the first space reactor to use thermionic energy conversion, in which electrons boil off a heated electrode and flow across a narrow gap to a cooler surface, thereby generating electric current. All earlier nuclear power sources in space have relied on less efficient but more reliable thermoelectric conversion, which transforms a temperature difference to a voltage at the junction of two dissimilar metals.

But accidental reentry is not the only danger that space nuclear power holds. Even those reactors that are launched or later boosted into a long-lived orbit present hazards because they could collide with orbital debris. Although it is unlikely at present, a collision between a nuclear reactor and one of the thousands of sizable objects traveling at a relative velocity of 10 kilometers per second could yield an abundance of radioactive fragments. Many of them would be driven into the lower orbits utilized by manned spacecraft and back into the earth's atmosphere within a few years. Unfortunately, most of the spent nuclear power supplies in orbit now reside in those parts of space near the earth that are most densely populated with debris.

Furthermore, even while they are operating safely, reactors can disrupt the operation of other satellites. To minimize mass and cost, orbiting reactors are largely unshielded. They thus produce strong emissions of radiation that can make it difficult for astronomical satellites to detect signals from distant sources. This phenomenon (which was kept secret by the U.S. government until 1988) has already significantly interfered with the work of orbiting gamma-ray detection systems such as that on board the National Aeronautics and Space Administration's *Solar Maximum Mission*.

The gamma rays emitted by orbiting reactors do not just outshine distant supernovas or black holes; in addition, the more energetic gamma rays interact with the outer shell of the reactor to produce streams of electrons and positrons. These charged particles are trapped in the earth's magnetic field, forming a temporary radiation belt. When another spacecraft passes through such a cloud, the positrons annihilate electrons in the spacecraft's skin, producing penetrating gamma rays that can overload the spacecraft's detectors.

These brief interruptions of astronomical observations afflicted the *Solar Maximum Mission* spacecraft an average of eight times a day for much of 1987 and early 1988, when the Topaz reactors were operating. Similar interference with the gamma-ray burst detector on board the Japanese *Ginga* satellite effectively blinded it during about a fifth of the same period.

NASA is endeavoring to limit the threat from orbiting reactors to its $500-million *Gamma Ray Observatory* mission, launched in April of this year. One proposed safeguard involves simply shutting off the gamma-ray burst trigger at times when it might be subject to interference. This strategy assumes, however, that only one or two low-power reactors, in predictable orbits, will be operating at any given time. If the number and operating power of orbiting reactors increase, the ability to conduct X- and gamma-ray observations from near-Earth platforms will be severely restricted.

Finally, orbiting reactors and RTGs pose not only environmental risks but political ones as well. From the first U.S. RTG-powered navigation and commu-

nications satellites through 20 years of *RORSAT*s, military requirements have often guided space nuclear power development, particularly when near-term civilian applications have been lacking. For instance, when the Strategic Defense Initiative was announced in 1983, the considerable power requirements of space-based missile defense appeared to provide a rationale for the SP-100, a proposed U.S. reactor that otherwise had no obvious application [*see box on page 20*].

So what good purpose is there for nuclear power in space? We believe that a useful distinction can be drawn between nuclear power in Earth orbit and in deep space. Many environmental hazards of space nuclear power vanish when an RTG or reactor is bound for another planet. Furthermore, deep-space applications are exclusively civilian and scientific or exploratory. (We include lunar missions in the deep-space category, even though the moon is, strictly speaking, in Earth orbit.)

During the past two decades, RTGs have made a major contribution to planetary exploration. The Mars probes, *Viking 1* and *2*, relied on nuclear power, as did the *Pioneer 10* and *11* and *Voyager 1* and *2* missions to the outer planets and the *Galileo* mission to Jupiter launched in October 1989. *Ulysses*, launched in October 1990 to explore the sun's polar regions, brought the U.S. total of nuclear-powered spacecraft to 25.

Indeed, the feasibility of many future space missions may depend on nuclear power sources. They are unaffected by their distance from the sun or by natural planetary radiation belts. Although solar panels or chemical power sources might hypothetically generate hundreds of kilowatts for a long period, the mass and volume of these systems would far exceed those of an equivalent reactor.

Nuclear power is essential, in particular, for missions to the outer planets. Beyond the orbit of Jupiter, the incident solar flux is less than 4 percent of the amount that reaches Earth, making solar power utterly impractical.

Eventually reactors not only may power spacecraft systems, they also may propel the craft. Various forms of nuclear propulsion are now under consideration as part of the U.S. Space Exploration Initiative. A reactor can heat a propellant directly or produce electricity to drive various kinds of thrusters. The U.S. conducted extensive ground-based tests of nuclear rockets in the 1960s, and both the U.S. and the Soviet Union have tested electric propulsion systems in space.

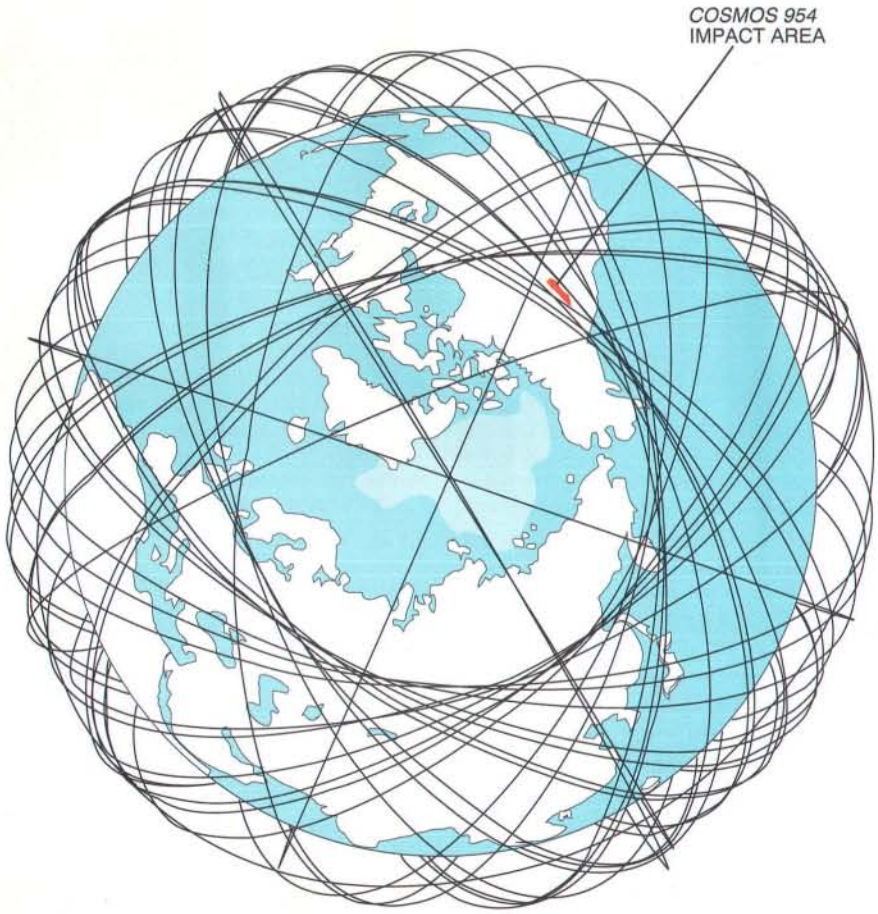
The availability of such advanced power and propulsion systems would increase the scientific return of future space missions. It would also reduce the need to wait years or decades for a favorable planetary alignment. These technologies would make possible such ambitious missions as the Thousand Astronomical Unit (TAU) program, for example, a 50-year voyage into nearby interstellar space. An electric propulsion system driven by a one-megawatt reactor could carry TAU far beyond the limits of our solar system, making possible unprecedented observations of the galaxy.

A nuclear reactor might be used to meet the power needs of a permanent lunar colony if one were ever established. Reactors weigh far less than the systems that would be needed to store solar energy during the 14-day lunar night. Alternatively, a base could be located at the north or south lunar pole, where there is continuous sunlight. One energy-intensive activity that would probably require a reactor is the production of propellant from lunar soil.

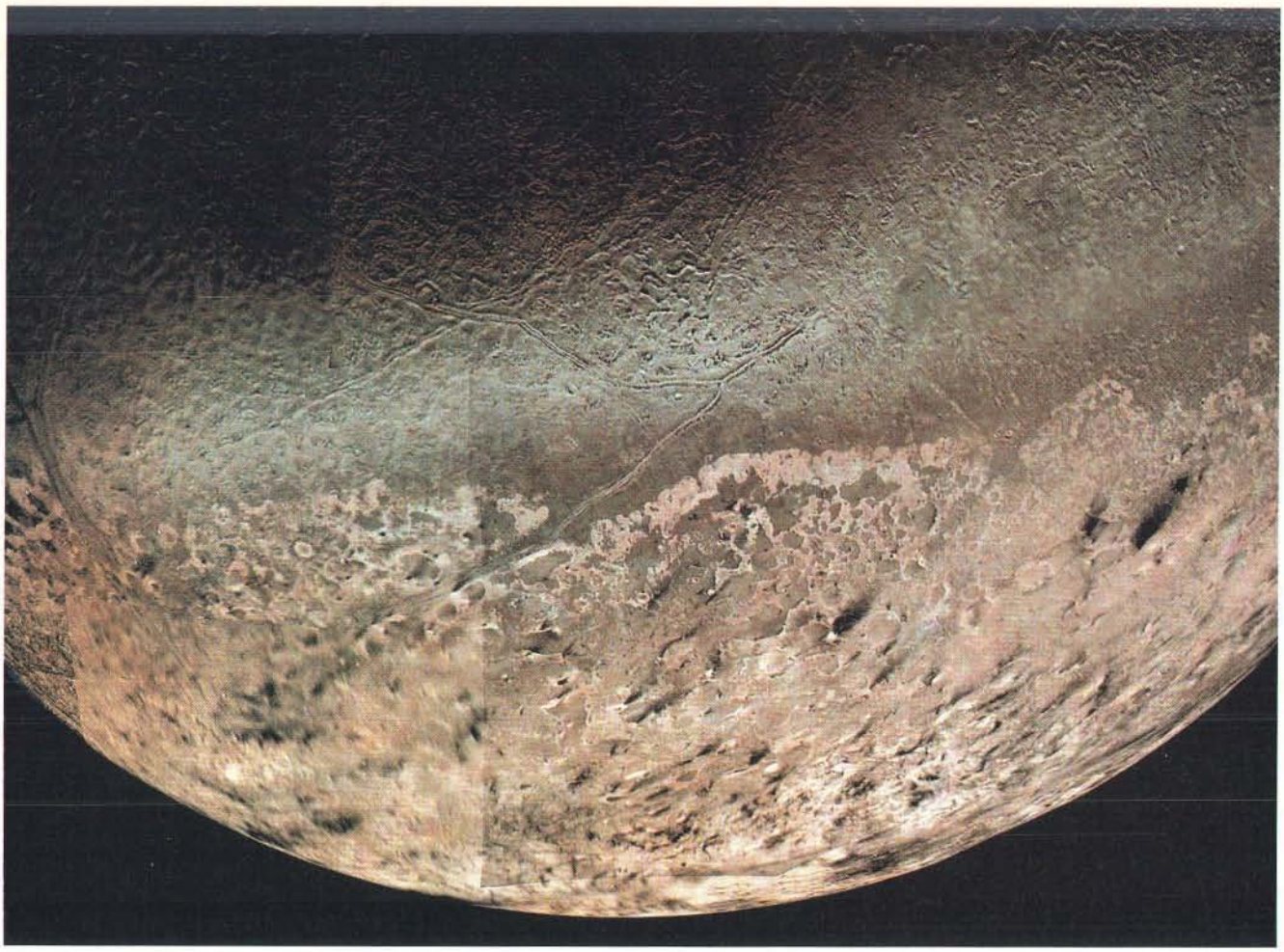
NASA is considering using a nuclear rocket to propel a manned mission to Mars. Doing so could perhaps reduce by half the mass that would have to be boosted into orbit from the earth's surface. Of course, the technical feasibility and the scientific value of such a mission are two distinct questions.

The development of any nuclear rocket would have to be carried out very carefully to minimize the environmental risk of an accident near the earth. Recent disclosures indicate that the U.S. Department of Defense is in fact working on such a rocket, but with imperfect attention to safety.

The highly classified project, codenamed Timberwind, is funded through the SDI. Timberwind was reportedly intended to boost massive space weapons into orbit on short notice, although the Defense Science Board concluded in 1990 that the nuclear rocket proj-



NUCLEAR-POWERED SPACECRAFT successfully deployed in Earth orbit by the U.S. and U.S.S.R. are estimated to number 42. (Two are in distant orbits not shown above.) All the spacecraft, or their jettisoned power supplies, are now in orbits high enough so that they will not reenter the earth's atmosphere until their radioactivity has substantially decayed. Most of them, however, orbit in a region populated with space debris. A collision could send a large number of radioactive fragments along trajectories that would reenter the atmosphere in a few years.



NEPTUNE'S MOON TRITON was photographed by the nuclear-powered *Voyager 2* spacecraft. The complex structure of Triton's surface has forced substantial revisions of theories of planetary geology. Radioactive power sources have been used on board several of the most scientifically productive space missions, including the *Viking*, *Pioneer* and *Voyager* probes.

ect was more likely to benefit deep-space propulsion. Preliminary testing of the fuel was carried out at Sandia National Laboratories in the late 1980s. The Timberwind reactor would operate at or near the melting temperature of its nuclear fuel, prompting concerns about the release of radioactive fission products.

SDI has proposed a suborbital flight test of Timberwind within the atmosphere near Antarctica. Such testing would violate accepted nuclear safety standards. Indeed, United Nations guidelines prohibit the operation of reactors on board spacecraft that have not achieved either a stable Earth orbit or an interplanetary trajectory.

In view of the various dangers they present, we favor an international agreement to ban nuclear reactors and RTGs from Earth orbit. This readily verifiable measure would eliminate many environmental hazards, enhance international stability and help to protect the space environment; it would also preserve the possibility of using nuclear power for scientific and exploratory missions in deep space.

Since space nuclear reactors possess a variety of distinguishing characteristics, an international agreement to prohibit them in orbit could be verified with confidence. In the first place, space reactors necessarily radiate large amounts of waste heat. They therefore give off a strong infrared signal that can be easily detected. Soviet *RORSATs* have been observed with a satellite-watching telescope at the Air Force Maui Optical Station on Mount Haleakala. An operating SP-100 reactor would be readily detectable at geosynchronous orbit or even beyond.

Operating reactors also emit strong gamma and neutron radiation signals that are easy to spot. Indeed, the type of gamma-ray interference that has disrupted scientific missions is uniquely produced by orbiting reactors and is a highly reliable, though unwelcome, sign of their presence.

It is true that a nation planning to "break out" of an arms-control treaty could conceivably place reactors in orbit without activating them, thus making the violation much more difficult to detect. But until large reactors have been thoroughly tested in space, they are unlikely to be covertly deployed.

Nuclear power has greatly enhanced space exploration. But it has also demonstrated the potential to produce significant environmental damage. Even if it is wisely controlled, nuclear power in space is likely to remain a challenging and costly technology.

FURTHER READING

ADVANCED POWER SOURCES FOR SPACE MISSIONS. National Research Council. National Academy Press, 1989.

GAMMA-RAY OBSERVATIONS OF ORBITING NUCLEAR REACTORS. Joel R. Primack in *Science*, Vol. 244, pages 407–408; April 28, 1989.

SPACE REACTOR ARMS CONTROL (special section). *Science and Global Security*, Vol. 1, Nos. 1–2; 1989.

The Quasar 3C 273

It is one of the most luminous objects in the known universe and the nucleus of an active galaxy. As astronomers scrutinize the spectrum of radiation from 3C 273, they are learning what makes quasars shine

by Thierry J.-L. Courvoisier and E. Ian Robson

The quasar 3C 273 lies about one fifth of the way from the earth to the edge of the known universe. Of all the objects in the cosmos, only a few other quasars surpass the energy and activity of 3C 273. On an average day, it is more luminous than 1,000 galaxies, each containing 100 billion stars. During one remarkable day in February 1988, the quasar erupted with a burst of radiation equivalent to lighting up stars the size of our sun at the rate of 10 million per second.

By monitoring 3C 273 in all domains of the electromagnetic spectrum and by observing variations in its luminosity, astronomers have begun to understand quasars and the physical processes that power them.

Since quasars were first identified some 28 years ago, astronomers have come to realize that quasars are the cores of extremely active galaxies. Quasars are unmatched in luminosity and hence are the most distant objects that can be detected in the universe. One of the most important discoveries about quasars is that their luminosity can vary greatly over periods of less than a year. This variability led investigators to the conclusion that the tremendous energy of quasars is radiated from a region many times smaller than the cores of ordinary galaxies.

THIERRY J.-L. COURVOISIER and E. IAN ROBSON have collaborated on many projects to investigate the quasar 3C 273 and other active galactic nuclei. Courvoisier is a research associate at the Geneva Observatory in Switzerland. After receiving his Ph.D. from the University of Zurich in 1980, he worked for the European Space Agency. Robson heads the department of physics and astronomy at Lancashire Polytechnic in England. In 1973 he earned his Ph.D. from Queen Mary College, where he was a member of the faculty until 1978. He then moved to Lancashire Polytechnic and two years later became the director of its observatories.

Quasars are powered by the gravitational energy that is released as gas and dust fall toward their massive, dense centers. Some of this energy channels particles into beams, blasting material out into the host galaxy at speeds close to that of light. Most of the energy is converted into radiation by a wide range of physical processes, probably occurring at different distances from the core. Yet quasars exhibit many phenomena that cannot be explained, and they remain one of the most puzzling objects in the universe.

On the whole, we know more about 3C 273 than any other quasar. It possesses a very wide range of properties, not all of which are shared by all quasars. The wealth of activity displayed by 3C 273, however, is a key in helping astronomers understand the phenomena at work in quasars.

The task of observing 3C 273 is as challenging as it is rewarding. After traveling through space for more than a billion years, only a tiny fraction of the radiation from 3C 273 reaches the earth. Capturing this radiation requires frequent observations using a battery of ground-based telescopes and satellite-borne instruments.

The effort began more than a century ago. The object known today as 3C 273 was first recorded on photographic plates as astronomers surveyed the stars in the constellation Virgo. It looked like nothing more than an ordinary, moderately bright star. Then in 1962 Cyril Hazard and his colleagues at Sydney University discovered that the starlike object occupied the same position in the sky as a strong source of radio waves. The radio emitter had been previously identified as 3C 273, which stood for number 273 in the Third Cambridge Catalogue of Radio Sources. Objects such as 3C 273 were subsequently described as quasistellar radio sources, or quasars.

In 1963 Maarten Schmidt of the Mount Wilson and Palomar Observatories deduced that the quasar 3C 273 was about three billion light-years away from the earth. The implications of this discovery were extraordinary. The quasar was by far the most luminous and distant object ever observed. Soon a few other quasars were identified that seemed to be even farther and brighter than 3C 273. At the time, many of Schmidt's colleagues had good reason to question these results. Yet as modern astronomers review the evidence collected during the past 28 years, we find little room to doubt that Schmidt was right.

Today astronomers are confident that quasars are not an isolated phenomenon in the universe. Quasars are the most active kind of celestial object that can be found in the nucleus, or center, of galaxies. To appreciate the properties of quasars, one should compare them with those of normal galaxies.

Galaxies, like our own Milky Way, are large assemblies of about 100 billion stars. In most galaxies the radiation observed on the earth comes mainly from the constituent stars, and the luminosity of the galaxy is therefore just the sum of the luminosities of the individual stars. In some galaxies a portion of the radiation originates in interstellar gases illuminated by hot stars. In yet another group of galaxies the nucleus generates most of the radiation and can even outshine all the associated stars. These are the so-called active galaxies, of which quasars are the extreme case.

The spectrum of radiation from galaxies—indeed, from any hot object—can exhibit three kinds of features: continuum, absorption lines and emission lines. Continuum radiation is composed of photons of all wavelengths. Its intensity changes smoothly from long wavelengths to short ones. At the wavelength of an absorption line, the intensity of the radiation is significantly less than the associated continuum radiation. An absorption line is created

as intervening gases absorb continuum emission at a particular wavelength. At the wavelength of an emission line, the radiation intensity is significantly greater than the continuum emission. These spectral lines are generated, for example, when interstellar gases absorb continuum emission and then reemit the absorbed energy as radiation of specific wavelengths.

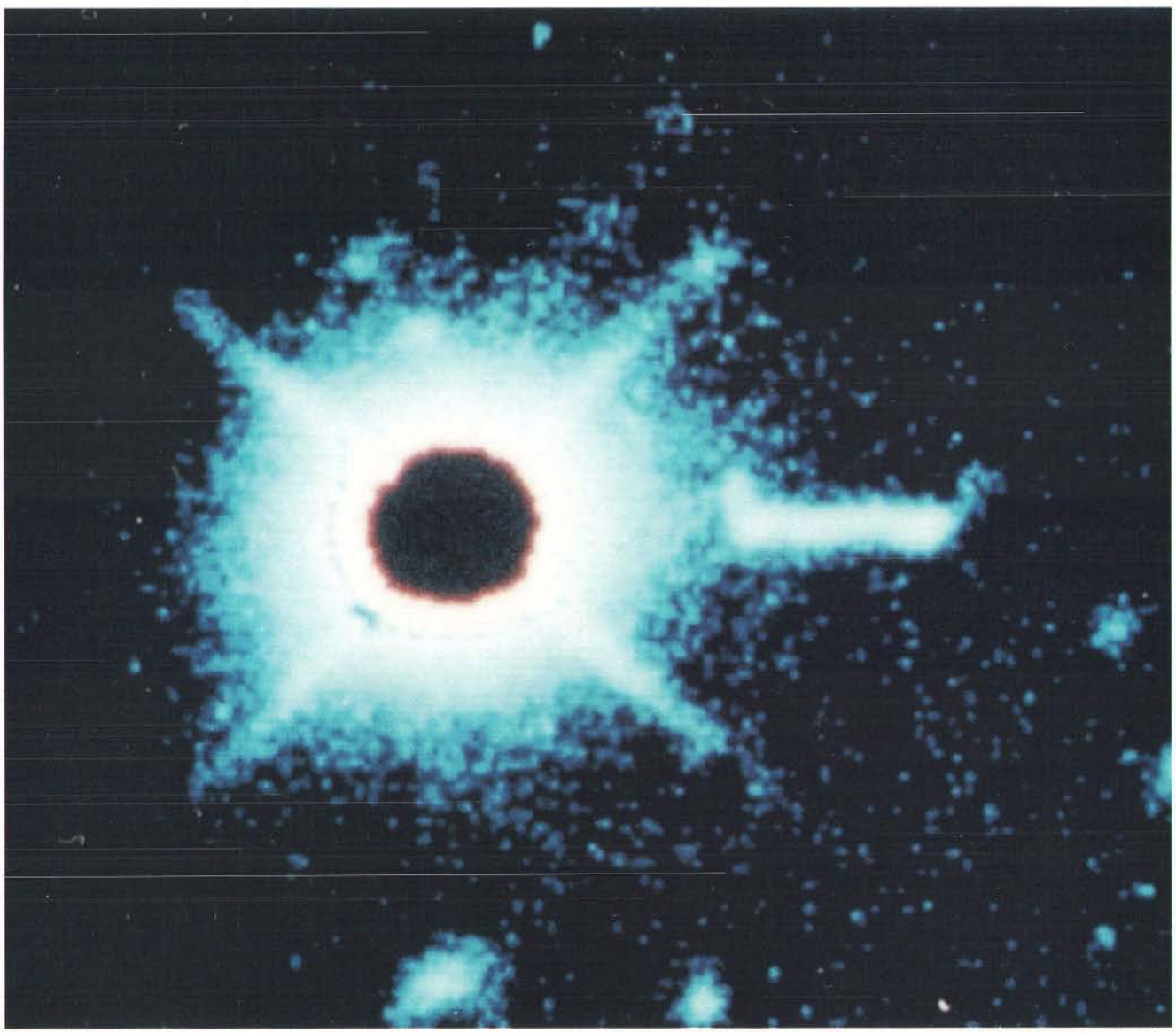
The spectrum of a normal galaxy consists mainly of continuum emission that is strongest in the visible domain. In contrast, the continuum emission from active galaxies and quasars is very intense from the infrared domain through the X-ray domain and is often strong at radio wavelengths. Spectra of normal galaxies typically display absorption lines but not emission lines. Active galaxies and quasars, however, exhibit strong emission lines in the visible and ultraviolet domains.

In active galaxies, emission lines are formed when ultraviolet and X-ray continuum emission strongly illuminate gas clouds. The atoms of the gas absorb the energy from the ultraviolet light and the X rays and then emit radiation at precisely defined wavelengths. The emission lines can indicate, therefore, that the active galaxy contains a strong ultraviolet and X-ray continuum source close to clouds of gas.

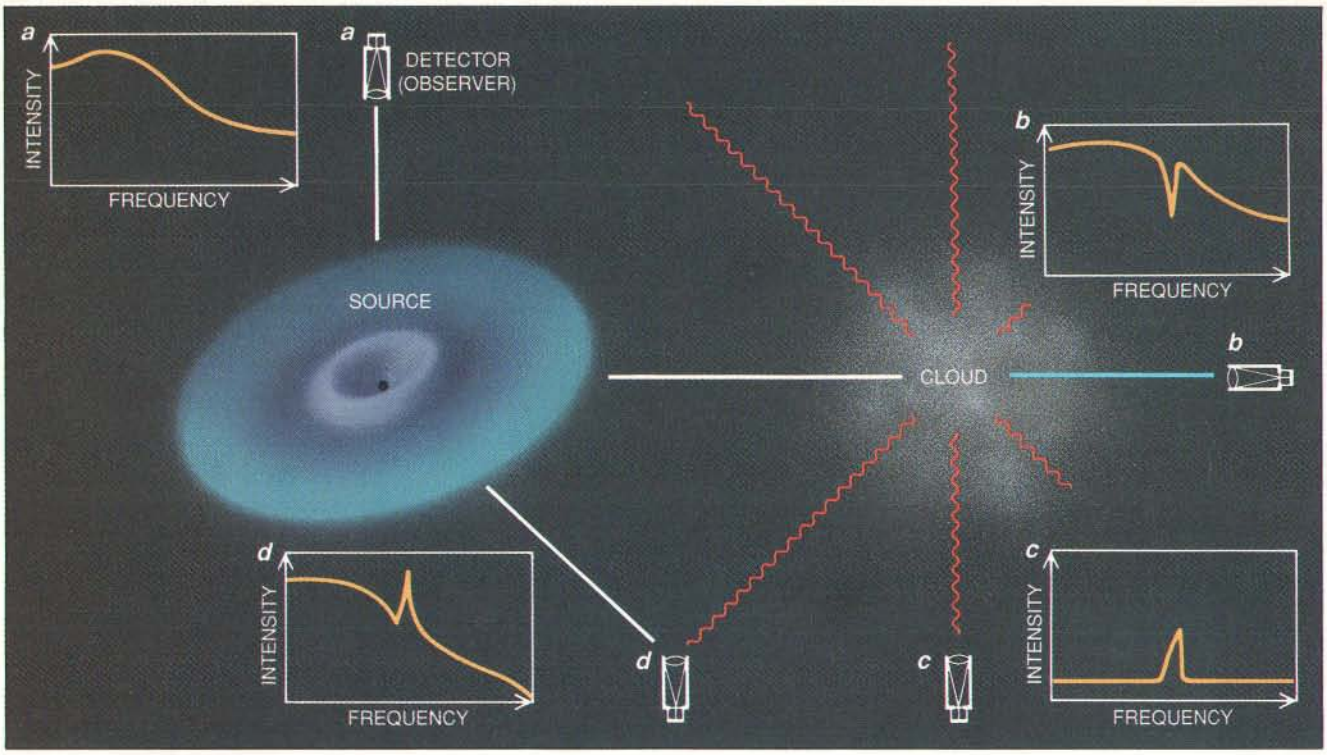
Even the center of normal galaxies can exhibit some properties of the more active ones, albeit on a much reduced scale. It is possible that the phenomena seen in quasars occur in a very much weakened form in the center of our own galaxy. It is also possible that many normal galaxies were much more active in the past and that the weak activity observed today is the relic of a dead or dormant quasar.

By studying the emission lines in quasars and active galaxies, astronomers can derive valuable information about the size, structure and dynamics of quasars.

Some emission lines observed in the spectra of quasars are broad rather than the sharp peaks generated in gases under laboratory conditions. This broadening implies that the gas clouds in the quasar are moving violently. By measuring the broadening, one can determine that the clouds are typically traveling at several thousand kilo-



JET OF QUASAR 3C 273 appears as a faint projection pointing toward the bottom of the photograph. (The other faint lines, which form an X, are artifacts.) To enhance the image of the jet, astronomers masked the bright light from the core.



SPECTRAL FEATURES are apparent when a source of continuum radiation and a cloud of gas are viewed in different ways. An observer very near the source sees a spectrum of continuum radiation (*a*). Another observer, looking through the cloud, measures an absorption line spectrum (*b*), which is produced because the cloud absorbs radiation at certain wavelengths. The light absorbed by the cloud is reemitted, and a third observer, looking at the cloud, sees an emission line spectrum (*c*). A fourth, viewing both the cloud and the source, sees the emission line superimposed on the continuum radiation (*d*).

meters per second. The high velocities most probably imply the existence of a strong gravitational force caused by a very massive object.

The velocity of the gases gives an indication of the mass associated with the central source, just as the velocity of the planets is linked to the mass of the sun. Unfortunately, to calculate the mass of the central source with confidence, one must know the distance of the gas from the center. This distance depends, however, on other parameters that are still uncertain. Nevertheless, astronomers can estimate the mass of the core of 3C 273. If the velocity of a typical cloud is assumed to be 5,000 kilometers per second at a distance of 10 light-years from the center, the core of 3C 273 can be calculated to be two billion times more massive than the sun.

One of the greatest controversies that surrounded quasars was just how distant they are from the earth. The distance from the earth to a galaxy is related to the speed at which that galaxy appears to be receding from the earth. This relation, known as Hubble's law, is a consequence of the cosmological expansion of the universe. When one observes the light emanating from distant galaxies, one is also looking into the past because of the time it took light to travel from the galaxy to the earth. These distant galaxies are seen at a time when the universe was younger, smaller and expanding more rapidly than at the present time. Therefore, the more distant the galaxy, the faster it is receding.

Because galaxies move away from the earth, the spectral lines that they emit are shifted. For instance, if a spectral line from 3C 273 is observed from the earth, its wavelength is 1.158 times the wavelength of the corresponding spectral line measured in the laboratory. This effect is known as the redshift. Because 3C 273 has a large redshift, one can infer that it is moving away from the earth at great speed. This high recession velocity then implies that the quasar is very distant.

One possible flaw in this argument—and a source of controversy—is that a galaxy's redshift can be altered by effects related to the great mass and energy of its nucleus. This led some workers to claim that the observed redshifts were caused, in part, by gravitational effects and other phenomena.

If such effects caused large redshifts in quasars, one would expect that the redshift associated with the nucleus of a galaxy would be greater than the redshift of the fringes of the galaxy. In all cases where both the galaxy and its active nucleus can be observed, however, the redshift of the active nucleus has been found to coincide with the redshift of the associated galaxy. This result gives great confidence that in cases where the active nucleus or quasar is too bright to allow observations of the underlying galaxy, the distance determined from the redshift of the emission lines provides the distance to the object. This argument and several others have convinced most astronomers that quasars are indeed at the distance indicated by their redshift.

The distance to 3C 273 can therefore be derived from Hubble's law. The distance is equal to the recession velocity divided by Hubble's constant. The recession velocity, which can be derived from the redshift, is 44,700 kilometers per second. Estimates for the Hubble constant range from 15 to 30 kilometers per second per million light-years. The distance to 3C 273 is therefore between 1.5 and three billion light-years.

By performing another calculation that takes into account the distance from 3C 273 to the earth and the observed flux of radiation, astronomers find that the luminosity of the quasar exceeds 10^{14} times that of the sun. This is about 1,000 times the luminosity of a typical normal galaxy.

Quasars radiate in almost all domains of the electromagnetic spectrum: radio, infrared, visible, ultraviolet, X and gamma. Most quasars are relatively weak in the radio domain, but 3C 273 emits roughly the same energy in all domains from radio to gamma regimes.

The flux of radiation from 3C 273 and other quasars can vary dramatically during periods of less than a year. The flux can also change at different rates in different regions of the spectrum. This variability was one of the very first discoveries about quasars and is an important clue to their behavior.

The spectrum of a quasar can be broken down into several emission components, each representing a portion whose intensity changes at the same rate. Each component probably arises from a different physical process occurring in the quasar. By observing 3C 273 for several years, we and our collaborators have been able to identify many distinct emission components. Some of them were already known, whereas others were revealed as a result of these studies.

The quasar 3C 273 can be observed in the visible domain from December until July, when it comes so close to the sun on the celestial sphere that our star effectively drowns out the object's light. Longer-wavelength radiations can, however, be detected for a more extended period.

Using a radio telescope in Finland, we observed the longest-wavelength component, which appears as a series of bumps in the radio domain [*see illustration at right*]. These features are thought to come from clouds of electrons that move through a strong magnetic field at velocities close to that of light. The magnetic field acts on the electrons in such a way that they follow curved paths and emit radiation. This process is called synchrotron emission. Astronomers believe that the synchrotron process is responsible for emissions in the radio and millimeter domains. To observe 3C 273 at wavelengths in the millimeter range, we used the James Clerk Maxwell Telescope in Hawaii and the Swedish Submillimeter Telescope at the European Southern Observatory in La Silla, Chile.

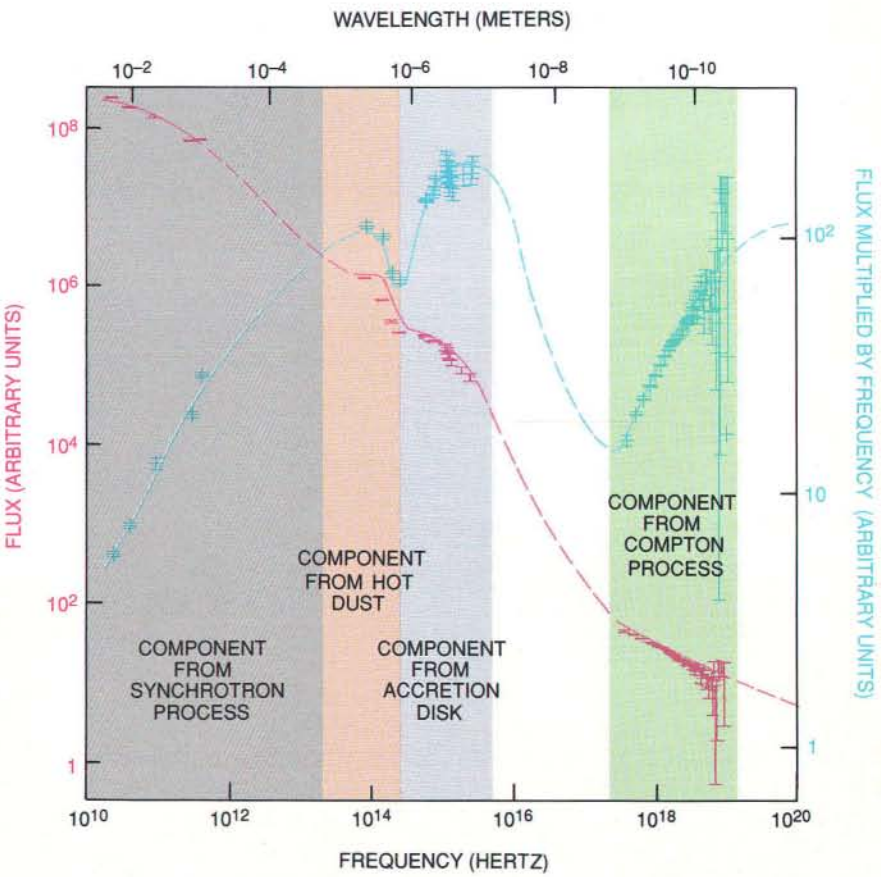
In 1986 we studied the variability of the infrared flux using the United Kingdom Infrared Telescope in Hawaii and instruments at the ESO in La Silla. We discovered that the synchrotron emission could not contribute significantly to the flux observed in the near infrared. This discovery agreed with previous puzzling observations that showed that the infrared emission was unpolarized. Synchrotron emission, however, is strongly unpolarized.

The origin of the infrared component is not yet known with certainty. One possible explanation is that the emission comes from cosmic dust heated to a temperature of about 1,500 kelvins. If this hypothesis is correct, the dusty region must be very large. The near infrared flux should therefore not change significantly over time, a prediction well corroborated by our ongoing observations.

Using the Swiss optical telescope in La Silla, an optical telescope in the Canary Islands and the *International Ultraviolet Explorer* satellite, we monitored a large, flat component known as the big blue bump. It is probably generated by the emission from the surface of a very hot object.

At present, astronomers cannot view the spectrum of 3C 273 between the ultraviolet and X-ray domains, for two reasons. Instruments that capture radiation in the extreme ultraviolet range are not nearly so sensitive as telescopes that monitor radio waves or light. In addition, radiation in the extreme ultraviolet is readily absorbed by interstellar matter, and so the flux of radiation that reaches the earth is small compared with the flux at other wavelengths. In fact, a new space telescope, the *Roentgen Observatory Satellite* (*ROSAT*), is just beginning observations in this spectral domain.

In the X- and gamma-ray domains, the flux decreases as a power law with decreasing wavelength. The X-ray component is probably produced by a phenomenon called inverse Compton emission. The phenomenon involves a pop-



FLUX OF RADIATION that reaches the earth from the quasar 3C 273 was measured at a variety of wavelengths (*red line*). This energy spectrum can be divided into four components, each arising from a different physical process. When each value of the flux is multiplied by its corresponding frequency, the blue line results. The peaks in this line represent regions in which the power output is greatest. All the observations were made in July 1987, except those in the millimeter-wavelength domain. Vertical lines show the range of uncertainty for each measurement.

ulation of high-energy electrons within the quasar. Such electrons scatter off photons to which they impart some of their own energy. In the process, the low-energy photons are transformed into X rays.

The *European X-ray Observatory Satellite* (*EXOSAT*) allowed us to view most of the X-ray domain, although technical obstacles make it difficult to monitor short-wavelength, or "hard," X rays. Further observations of X rays were made using the Japanese satellite *Ginga*.

To determine how much each emission component contributes to the total luminosity of the quasar, astronomers can use a simple calculation: the luminosity at a particular wavelength is related to the wavelength multiplied by the flux density at that wavelength as measured on the earth. Such calculations show that all the emission components have comparable luminosities, a surprising fact considering that each component seems to originate from a different physical process. Two components, however, are dominant: the extreme ultraviolet and the hard X rays. By a quirk of fate, these two regions are the most difficult to observe, leading to some uncertainty about the total energy emitted by quasars.

The time scale of variation of any object is crucial because it can indicate the approximate size of the source of the radiation. As a simple example, consider a line of 10 light bulbs. If one wishes to decrease the total luminosity significantly, a large number of the light bulbs must be switched off, say, at least six. To do this, one must send a signal that instructs the bulbs to turn off. In the physical world, no signal can travel faster than the speed of light. The dimming process will therefore take at least the time light needs to cross the distance from the center of the line of bulbs to the most distant bulb to be turned off. Likewise, the dimming process for astronomical objects must take at least the time light needs to cross the emission region. Hence, an upper limit on the dimensions of the source can be determined by multiplying the speed of light by the time scale of variation.

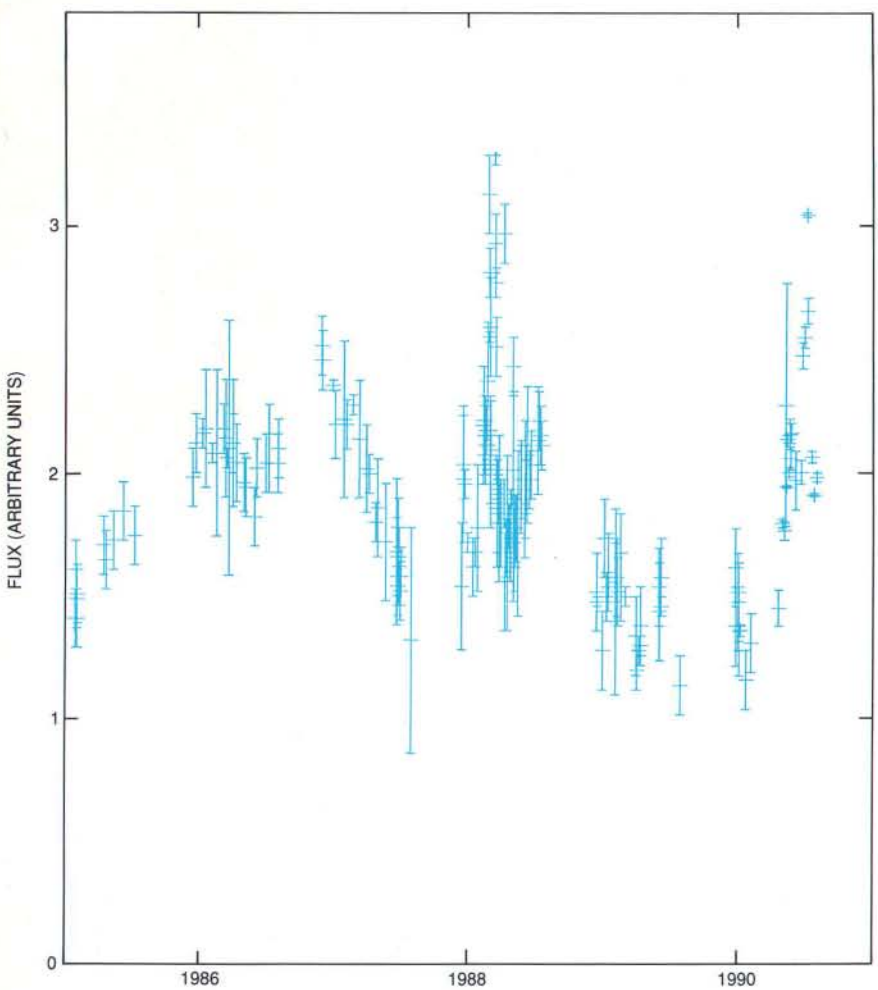
Photographs of 3C 273 taken during the past century show that the object's visible flux varies over many different time scales, most longer than about 10 days. We and our colleagues have also recently observed that most spectral domains vary over time scales of about one to a few months. This variation indicates that the diameter of the source of continuum emission must be much less than one light-year, less than the distance from the solar system to the nearest star.

We were surprised to discover that the variations of the different emission components are not correlated. This apparent lack of correlation revealed that the structure of 3C 273 was much more complex than expected. This complexity is one of the reasons astronomers have found it very difficult to untangle the relations between the geometry of the emission components and their physical origins.

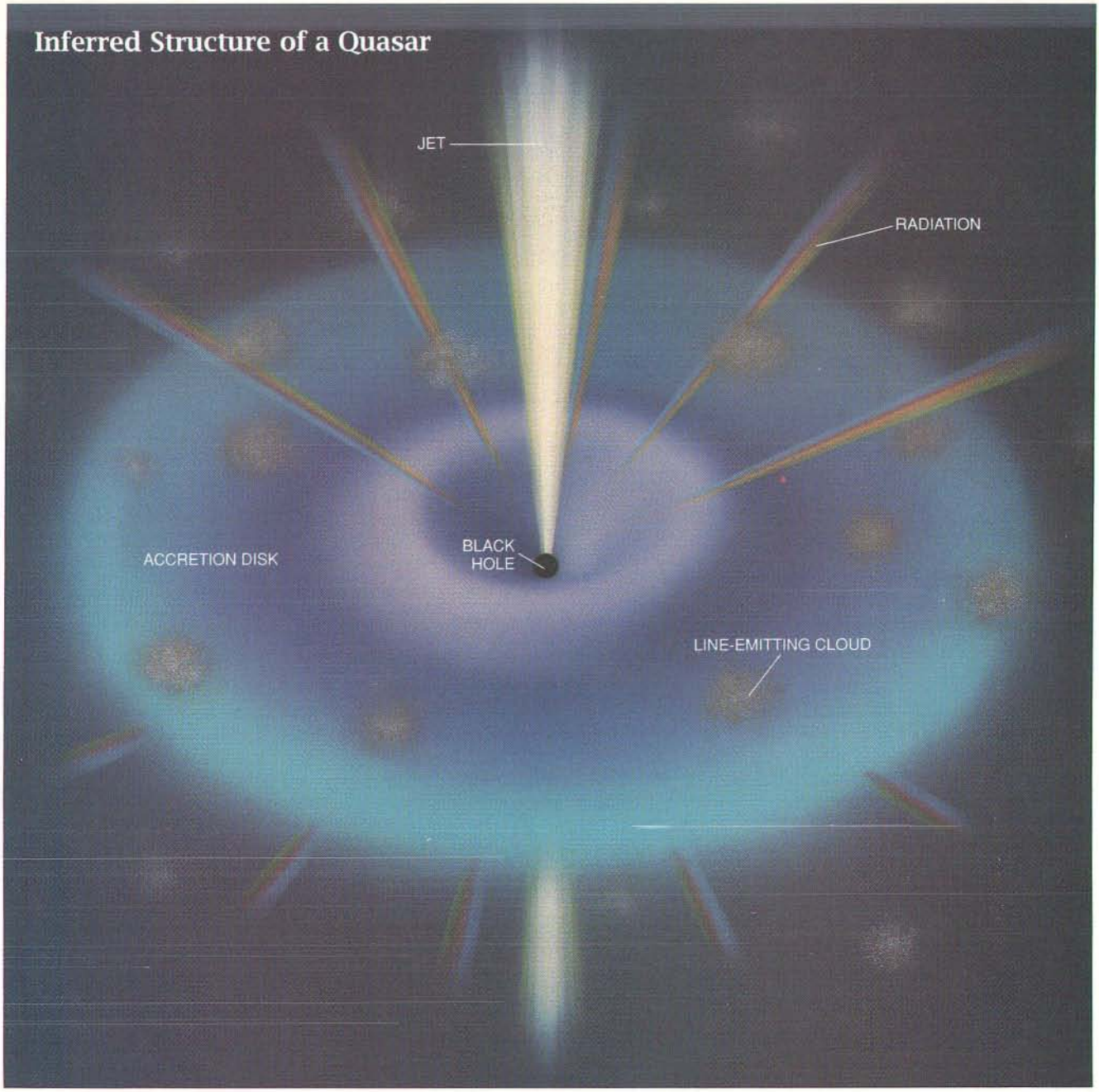
During the 1988 observing season we and our colleagues made an unusual discovery concerning such variations. In January we noted that 3C 273 was in a state of low activity. Suddenly in February, it flared up rapidly and repeatedly. The visible and infrared flux changed by as much as 50 percent. The time scale of these changes was unprecedented: the fastest variations happened in the course of a day. This behavior was totally unexpected for an object as luminous as 3C 273 (although rapid flaring in the X-ray domain had been seen before in some low-luminosity nuclei of galaxies).

The violent activity lasted for four months, during which time the activity in the optical spectrum peaked five times. The peaks of maximum flux were separated on average by about 15 days. Indeed, two of the maxima were only two days apart. The fastest variation in flux was a decrease of about 15 percent in 24 hours. This rapid decline was equivalent to the turning off of some 10 million suns per second!

Repeated flares were also recorded in the infrared domain, but only two maxima were caught, because the infrared component was observed less frequently than was the optical domain. The two maxima coincided with two of the visible maxima. One infrared maximum was roughly twice the "quiescent" emission. The fastest change observed in



FLUX observed from the quasar 3C 273 varies considerably over time in the visible domain. In February 1988 the flux changed rapidly and dramatically, indicating that the quasar was in a state of violent activity. Vertical lines show the range of uncertainty for each measurement.



the infrared was an increase of about 40 percent in 24 hours. This rate of change corresponds to the switching on of about 10 million suns every second.

This very active period came at a most opportune time: when 3C 273 was clear of the sun and our collaboration was ready to react quickly. We managed to cover the event thoroughly, sometimes making daily observations. Yet the flaring was so rapid that we suspect that many features went undetected.

In May, 3C 273 settled back to its normal state for the rest of the observing season. We and our colleagues are carefully analyzing the periods before and after the 1988 maxima in an attempt to understand why 3C 273 suddenly erupted with activity. The 1988 observations were the first records of such violent activity in quasars, showing either that this state is relatively rare or that astronomers have been unlucky not to have caught other such events in the past. To determine just how rare these events are, we will need to monitor 3C 273 and other quasars for many years to come.

Fast variability often indicates that regions of the source are moving at velocities close to that of light (relativistic speeds). One must first know something about the behavior of high-energy electrons and the concept of brightness temperature before understanding the phenomenon.

High-energy electrons in a very compact source cool as they scatter off infrared photons produced by synchrotron emission. The electrons give some energy to the photons and shift them to the X-ray domain. This combined process, called synchrotron self-Compton emission, is related to the size of the source. Specifically, as the size of the source decreases, the ratio of the self-Compton component to the synchrotron component increases.

The luminosity and the size of an

emitting region can be expressed as a quantity called the brightness temperature. This quantity is not the actual temperature of the object. It is simply a convenient way to express the luminosity and size. An object that exhibits synchrotron self-Compton emission will not exceed a brightness temperature of 10^{12} kelvins by very much or for very long. At temperatures above 10^{12} kelvins, radiation is largely released through self-Compton emission, which cools the object back to 10^{12} kelvins. In addition, an object whose brightness temperature exceeds 10^{12} kelvins would emit most of its flux in the X-ray domain and not in the infrared domain. Hence, 10^{12} kelvins is essentially the highest possible brightness temperature. It is technically known as the Compton limit.

In some cases, however, the observed brightness temperature can exceed 10^{12} kelvins. How can this be? The simplest interpretation is that the source of emission is moving at relativistic speeds toward the observer, and as a result, the observed radiation is concentrated into a beam and is strongly intensified as compared with radiation from an identical source at rest. This apparent change is called beaming.

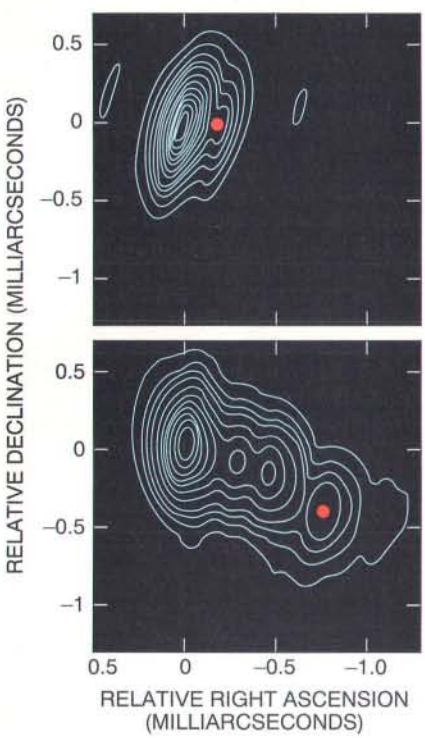
The measured flux of the source is therefore greater than what would be detected if the source were moving at nonrelativistic speeds. Hence, the brightness temperature recorded from a relativistic source is greater than the temperature that would be measured from a nonrelativistic source. In this way, the observed brightness temperature can be greater than the Compton limit of 10^{12} kelvins.

The very fast variations in the optical and infrared flux of 3C 273 imply a brightness temperature still well below the Compton limit. Observations at wavelengths longer than infrared suggest, however, that the brightness temperature produced during the February and March events of 1988 were well in excess of the Compton limit. Hence, beaming is suspected. Further evidence comes from observations that the radiation at millimeter wavelengths varies rapidly (on a time scale of several hours). This rapid variation implies extremely high brightness temperatures, greatly exceeding the Compton limit. Apparently, the source of the flare is moving at relativistic speeds toward the earth.

Another powerful line of argument points to the existence of relativistic motions in 3C 273. Astronomers can obtain very high resolution radio maps using a technique known as very long baseline interferometry (VLBI) [see "The Very-Long-Baseline Array," by Kenneth I. Kellermann and A. Richard Thompson; SCIENTIFIC AMERICAN, January 1988]. The VLBI network produced radio images of several quasars, revealing in most cases compact cores and jets formed of successive "blobs" of synchrotron emission.

Perhaps the most remarkable aspect of these blobs in 3C 273 and other quasars is that they appear to move away from the core at velocities several times that of light. The theory of special relativity holds that physical speeds do not exceed the velocity of light anywhere in the universe. But if an object ejects material in the observer's direction at velocities slightly less than the speed of light, the material can appear to be moving at speeds exceeding the fundamental limit [*see box on opposite page*].

Recently workers found that a new blob appeared at the same time as did the rapid variations. It so happened that the VLBI network observed 3C 273 in June 1988 and in March 1989. The 1988 observation yielded an image of the source in which a new blob had appeared [*see illustration on this page*]. The 1989 image revealed that the blob had moved some distance from the core. Using these two measurements, investigators deduced the apparent velocity of the blob and therefore the time at which it had been ejected from the core. They found that the new component emerged during the period when 3C 273 was releasing bursts of radiation in the optical and infrared domain. So it seems that the violent activity is closely related to the birth of components in the jet. We hope that this connection will help us understand how the jets are formed.



RADIO IMAGES obtained on June 25, 1988 (*top*), and March 9, 1989 (*bottom*), show a blob of radiation (*red dot*) moving away from the core of the quasar 3C 273. From these images, one can deduce the velocity of the blob and its time of birth. The blob seems to have been generated during the quasar's violent period in February 1988. The illustrations were based on data from T. Krichbaum of the Max Planck Institute for Radioastronomy in Bonn.

Successful theories of quasar activity must explain the presence of relativistic jets as well as the broad and narrow emission lines and the several continuum emission components. As yet, no theory exists that is this complete. Current theory suggests that high-energy continuum emission is radiated from the center of the quasar and illuminates clouds of gas some distance away. Astronomers have some understanding of the processes at the origin of the continuum emission in terms of synchrotron emission, Compton emission and so on. None of this, however, tells us where the enormous quantity of energy needed to fuel these processes is generated. In other words, although we know the radiation is produced by very energetic electrons, we do not as yet know how the electrons acquire their energy.

Workers therefore confront the difficult task of explaining how an object only a few light-years across or even less can emit as much radiation as 1,000 galaxies.

Most investigators now believe that quasars are ultimately powered by the gravitational energy associated with very massive and dense objects known as black holes. The evidence is circumstantial. Some clues come from the rapid variations, which show that the central source is very compact. Other clues are mass estimates deduced from the motion of gas clouds.

Astronomers can also point to the fact that it is possible to extract much more energy from gravitational forces than from any other kind. They have come to this conclusion, in part, by studying gravitational and nuclear processes in X-ray "bursters" in our galaxy. The cores of these objects are extremely dense, hot bodies known as neutron stars. As material falls onto the surface of the star, gravitational energy is released. This material then undergoes a nuclear-burning process that

produces bursts of X rays. The analysis of X-ray bursters shows that gravitational forces release 100 times more energy than nuclear forces do. It is therefore natural to expect that gravitational energy is also the primary source in quasars.

The emission from any compact object cannot be boundless. It is limited by the fact that radiation itself exerts a pressure. Extremely luminous sources can emit radiation so intense that it blows surrounding material away. If an object gains energy as matter is attracted toward it, the resulting luminosity cannot exceed the Eddington limit, the luminosity at which the radiation pressure will begin to push matter away. As the mass of the object increases, so does the Eddington limit. Hence, if one assumes that the luminosity of an object is close to the Eddington limit, one can estimate the object's mass.

For the quasar 3C 273, the mass deduced from the luminosity is billions of times the mass of the sun. The mass estimated from the luminosity of 3C 273 and other quasars is not inconsistent with that derived from studies of motions of gases. Within the uncertainties it is therefore reasonable to think that gravitation plays a major part in the energy release of quasars.

Another parameter that determines the energy output of a quasar is the accretion rate: the amount of mass falling toward the gravitational center per unit time. Because quasars require a great amount of gravitational energy per unit time, they must also have high accretion rates, that is, a large quantity of mass must fall toward the gravitational center. Indeed, the accretion rate of 3C 273 can be deduced from its power output of about 10^{40} watts. If the quasar converts gravitational energy to radiative energy with an efficiency of about 10 percent, the accretion rate is about 10^{24} kilograms per second, or a few times the mass of the sun per year. This calculation suggests that every year the equivalent of a few stars like the sun falls into the gravitational center of 3C 273.

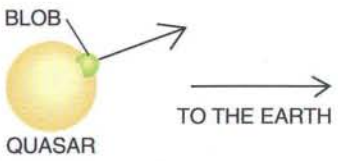
To emit so much energy, a quasar requires a core that is very compact as well as massive. The densest body known to physics is a black hole, an object whose gravitational forces are so great that neither material nor radiation can escape. According to the theory of general relativity, a black hole is the ultimate stable configuration for a very massive object. Many astronomers therefore believe that whatever the original state of the massive compact object in quasars, a black hole will be formed. Unfortunately, no direct observation has ever been made that could either confirm or refute the existence of black holes in quasars.

The matter surrounding a massive black hole is likely to spiral toward the center rather than taking a direct path. The matter will therefore form a disk around the central mass. This accretion disk is a structure that is seen around many compact objects in our galaxy. Accretion disks are so dense that radiation escapes only from the surface of the disk, just as stars emit light mainly from the surface region called the photosphere. Investigators have shown that the luminosity of the accretion disk's photosphere is possibly responsible for the big, blue bump observed in the visible and ultraviolet domains of the spectrum of 3C 273.

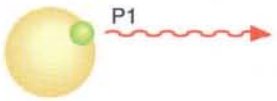
The structure and dynamics of quasars remain a mystery in many respects. Astronomers cannot yet figure out the geometries of the synchrotron component, the infrared component and the X-ray component. Nor can they deduce what relation these components have to the fueling mechanisms and the accretion phenomena. The observations we have made with our colleagues have shown that these issues are complex. These components might be closely associated with the generation of the relativistic motions. Although our knowledge of quasars is still far from complete, we hope that extensive observations and theoretical insights will continue to provide clues, like pieces of a puzzle, enabling the overall picture to be assembled.

How an Object Can Appear to Move Faster Than Light

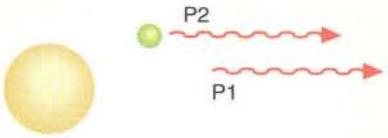
One of the most fundamental laws of physics is that all radiation and matter can move only as fast as the speed of light, that is, 300 Mm/sec (million meters per second). If a blob of matter is ejected from a quasar at speeds close to that of light, it can appear to move faster than light, however. This effect has a straightforward explanation. Consider a blob that is moving at 240 Mm/sec toward the earth and at 90 Mm/sec perpendicular to the line of sight. (Equivalently, one can say that the blob is moving at 256 Mm/sec at an angle of 20 degrees from the line of sight.)



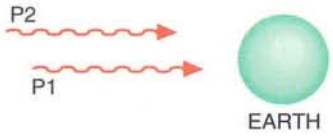
At the instant the blob leaves the core, it releases a photon of radiation, designated photon 1.



After one second, photon 1 has traveled 300 Mm, and the blob has moved 240 Mm from the quasar toward the earth. It then emits a second photon.



As the two photons travel through space, photon 2 will lag 60 Mm behind photon 1 in the direction parallel to the line of sight. The two photons will also be separated by 90 Mm in the direction perpendicular to the line of sight.



Photon 2 will arrive at the earth 0.2 second after photon 1 because the time delay equals the "parallel" change in distance (60 Mm) divided by the velocity of the photons (300 Mm/sec). The apparent velocity of the blob across the sky is the "perpendicular" change in distance (90 Mm) divided by the time delay (0.2 second). Hence, the blob appears to travel at 450 Mm/sec, or 50 percent faster than the speed of light.

FURTHER READING

THEORY OF EXTRAGALACTIC RADIO SOURCES. M. C. Begelman, R. D. Blandford and M. J. Rees in *Reviews of Modern Physics*, Vol. 56, No. 2, Part 1, pages 255–351; April 1984.

BLACK HOLE MODELS FOR ACTIVE GALACTIC NUCLEI. M. J. Rees in *Annual Review of Astronomy and Astrophysics*, Vol. 22, pages 471–506; 1984.

QUASAR ASTRONOMY. Daniel W. Weedman. Cambridge University Press, 1986.

ACTIVE GALACTIC NUCLEI. T. Courvoisier, R. D. Blandford, L. Woltjer, M. Mayor and H. Netzer. Springer-Verlag, 1991.

Streptococcal M Protein

The bacteria that cause strep throat and rheumatic fever depend on this cell-surface molecule to evade the body's defenses. The key to the protein's power is its remarkable structure

by Vincent A. Fischetti

Just as mammals have hair and birds have feathers, microorganisms have sophisticated structures on their exteriors that help to ensure their survival. Some of these structures are receptor molecules that direct a microbe to the specific environmental niche that is its natural habitat. Other molecules prevent an organism from being destroyed by natural processes. Pathogens that cause diseases in people are probably the microorganisms whose surface features have been best studied. Such viruses, parasites and bacteria are covered with protein or sugar molecules that help them gain entry into a human host by counteracting his or her defenses.

One such molecule is the M protein produced by certain streptococcal bacteria. Early studies of its structure made it seem unique, but the molecule now appears to embody a common motif shared by many bacterial surface proteins. The M protein can therefore serve as a model for other surface proteins, a fact that may hasten the development of some antibacterial therapies. A thorough understanding of the M protein has already led to its use as a possible candidate for a vaccine to control streptococcal infections.

VINCENT A. FISCHETTI is professor and co-chairman of the laboratory of bacterial pathogenesis and immunology at the Rockefeller University. He also serves as editor in chief of the journal *Infection and Immunity*. Fischetti was born and educated in the New York area. After receiving a bachelor's degree in bacteriology from Wagner College in 1962 and a master's degree in microbiology from Long Island University in 1967, he went on to earn his doctorate in microbiology from the New York University School of Medicine in 1970. Fischetti holds a MERIT Award from the National Institutes of Health for his work on the streptococcal M protein.

Streptococci are a group of bacteria with the capacity to grow in chains. Many varieties are part of the normal bacterial flora in humans and are not especially harmful. *Streptococcus salivarius*, for example, is found routinely on the tongue. *Streptococcus mutans* is commonly and exclusively associated with the teeth (and is usually involved with dental caries).

A particular subgroup of streptococcal bacteria, called group A and represented by *Streptococcus pyogenes*, is a human pathogen. Between 20 and 30 million cases of group A streptococcal infections occur every year in the U.S. alone. These cases include infections of the skin and throat, forms of pneumonia and a recently identified disease resembling toxic shock. The most common infection is acute streptococcal pharyngitis, or strep throat, which occurs predominantly in school-age children. Strep throat would qualify as a major worldwide health problem if judged only by the time lost from school and work and by the amount spent on related doctor's fees.

Strep throat's toll is much greater, however. In as many as 4 percent of the pharyngitis cases that are untreated or treated ineffectively, the strep infection leads to acute rheumatic fever, a disease that damages the heart, particularly the heart valves. Currently most surgical procedures to correct damaged heart valves in the U.S. are performed on patients who had rheumatic fever as children. Those heart-valve injuries are not direct effects of the bacteria; instead they result from an immunologic assault. Streptococci fool the immune system into attacking some of the body's own tissues.

Although rheumatic fever is not a serious problem in the U.S. today, it is still a major health hazard in developing nations. About 1 percent of the children between ages five and 15 in the developing world have rheumatic heart disease. By one estimate, for example, it affects nearly six million school-age children in India.

Group A streptococci can persist in tissues for weeks, primarily because of the M protein on their outer surface. The M protein gives a streptococcus the ability to resist ingestion by human phagocytes, the white blood cells that seek and destroy invading microorganisms. Rebecca Lancefield of the Rockefeller Institute (now University) discovered this fact about 60 years ago, and it is easily demonstrated by a simple experiment. If streptococci are placed in a drop of human blood, the phagocytes will avoid the bacteria with M proteins on their surface, but they will actively attack those that lack the protein.

Thus, it is apparent that group A streptococci have developed a system for avoiding some of the antimicrobial defenses of a human host. Yet resistance to an infection by these bacteria is possible if the host's body can produce antibodies directed against the M protein. Such antibodies will neutralize the protective capacity of the M protein and allow the streptococcus to be engulfed and destroyed by phagocytes.

Unfortunately, there are more than 80 different serotypes, or varieties, of M protein. Laboratory tests suggest that antibodies against one serotype do not offer protection against others. As a result, researchers believe that exposure to one serotype of group A streptococcus will not immunize an individual against all further streptococcal infections. The system that the streptococci have devised for changing the M molecule to avoid antibody recognition is called antigenic variation. It is not unique to streptococci. Other disease organisms, including viruses and parasites, commonly rely on antigenic variation to avoid immune recognition of their surface molecules.

My colleagues and I at the Rockefeller University, among many other

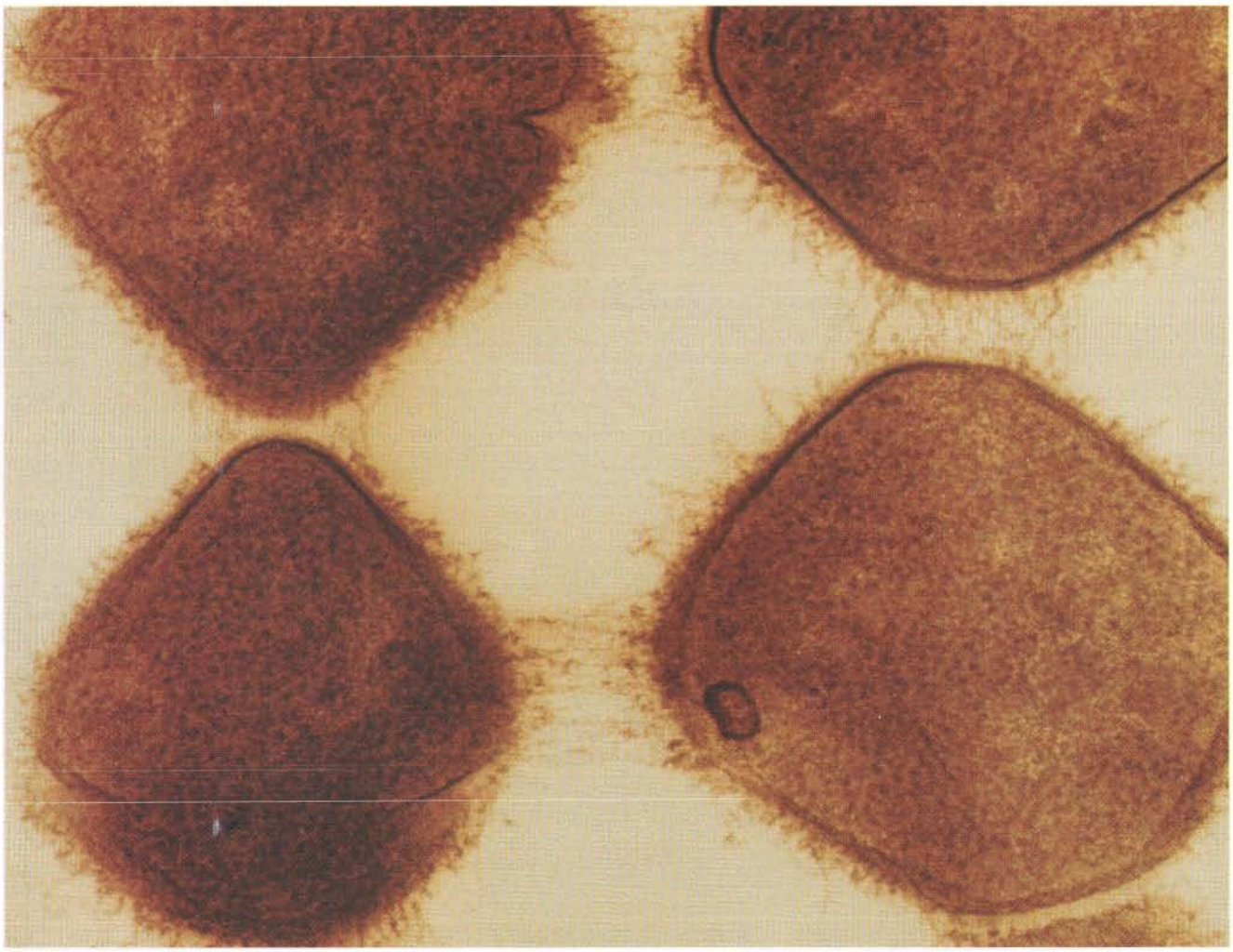
scientists, have worked toward understanding how the M molecule changes antigenically and how it helps the streptococcus evade phagocytosis (attack by phagocytes). The key to these characteristics seems to be in the M protein's structure, which helps the bacteria fend off the body's defenses in at least three ways. That knowledge has enabled us to develop strategies—and a potential vaccine—for controlling streptococcal infections. What we have discovered also gives us insights into the structure and function of surface proteins on other bacterial pathogens.

When we look at ultrathin sections of group A streptococci under an electron microscope, the M proteins appear as hairlike projections on the surface of the bacterial cell wall. The cell wall is a strong but pliable physical barrier that protects the more fragile cell membrane below it. That membrane controls the inward flow of nutrients and the outward flow of wastes and other products. It also bears many enzymes necessary for cellular metabolism and protein synthesis. In most bacteria, a layer of complex carbohydrates is linked to the outer surface of the cell wall.

Streptococci are categorized as gram-positive bacteria because they have relatively thick cell walls that retain a specific chemical stain after an identification procedure. (This staining technique was developed by the Danish bacteriologist H.C.J. Gram.) The streptococcal cell wall consists of long linear chains, or glycan units, of a disaccharide sugar molecule, *N*-acetylglucosamine-*N*-acetylmuramic acid. These chains are interconnected by short peptides, or amino acid chains. Because of its structure, the cell wall is often called a peptidoglycan. The peptidoglycan of a gram-positive organism may be envisioned as roughly 10 layers of glycan strands cross-linked by peptides to form a meshlike bag. The peptide constituent sometimes varies among different gram-positive species, but the fundamental structure of peptidoglycans is almost always similar.

An important first step in deciphering the structure of the M protein was learning the sequence of amino acids in one such molecule. In 1986 our laboratory, in collaboration with June R. Scott of Emory University and Susan Hollingshead, who is now at the University of Alabama, succeeded at this task. We cloned (isolated and copied) an M protein gene—specifically, the gene for the M protein of type 6 (M6) streptococci. Other investigators had previously determined portions of the amino acid sequences from M molecules. With a cloned gene, we could decipher the complete amino acid sequence and begin to study the M6 protein's whole structure.

Approximately 80 percent of the M6 molecule is made up of four dis-



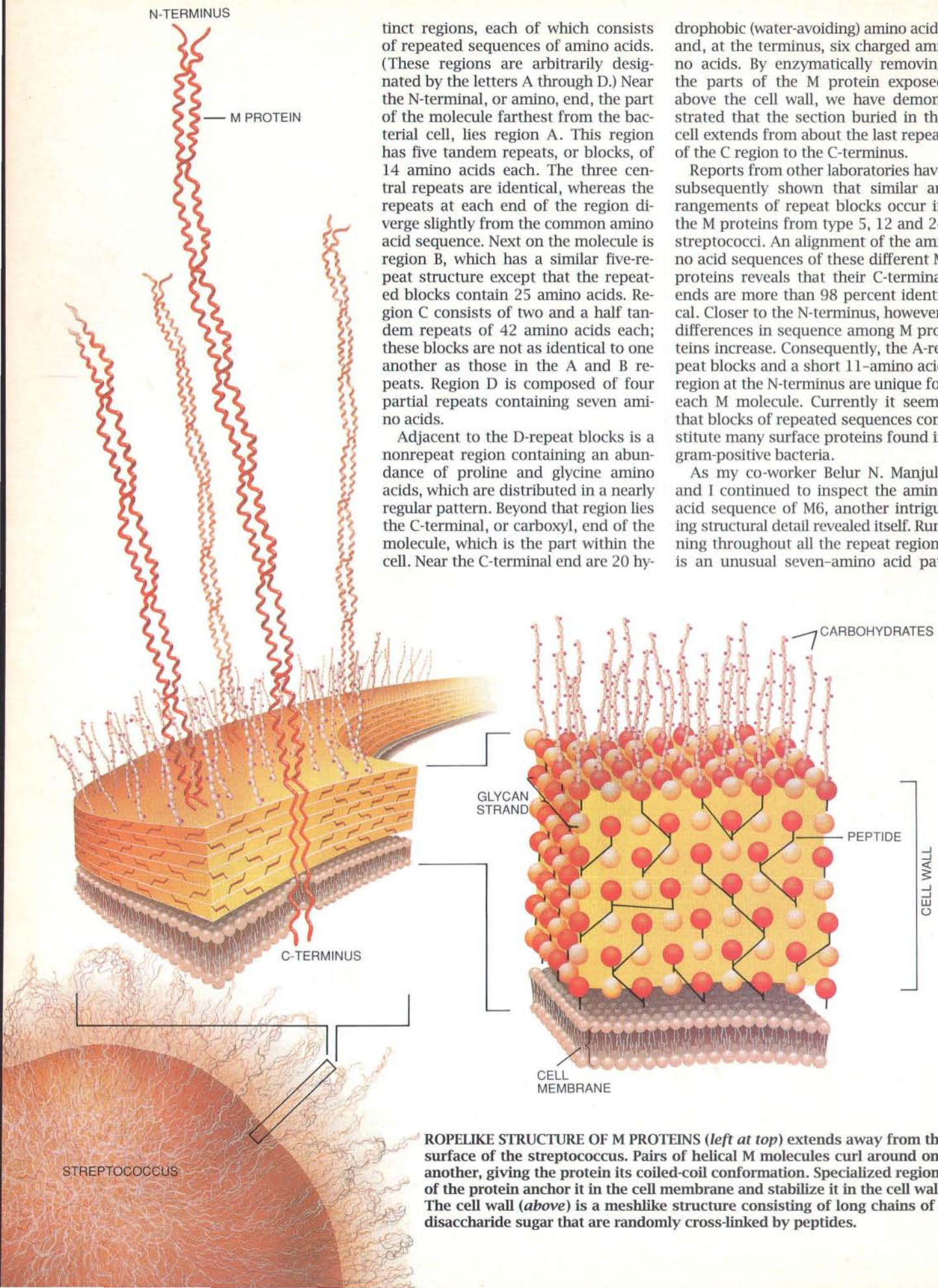
M PROTEIN appears as long, hairlike filaments on the surface of group A streptococci. It helps the bacteria evade their human hosts' immunologic defenses and avoid being cleared from the body. Recent studies of the streptococcal M protein have revealed how specific structural features of the molecule enable it to carry out these functions.

tinct regions, each of which consists of repeated sequences of amino acids. (These regions are arbitrarily designated by the letters A through D.) Near the N-terminal, or amino, end, the part of the molecule farthest from the bacterial cell, lies region A. This region has five tandem repeats, or blocks, of 14 amino acids each. The three central repeats are identical, whereas the repeats at each end of the region diverge slightly from the common amino acid sequence. Next on the molecule is region B, which has a similar five-repeat structure except that the repeated blocks contain 25 amino acids. Region C consists of two and a half tandem repeats of 42 amino acids each; these blocks are not as identical to one another as those in the A and B repeats. Region D is composed of four partial repeats containing seven amino acids.

Adjacent to the D-repeat blocks is a nonrepeat region containing an abundance of proline and glycine amino acids, which are distributed in a nearly regular pattern. Beyond that region lies the C-terminal, or carboxyl, end of the molecule, which is the part within the cell. Near the C-terminal end are 20 hydrophobic (water-avoiding) amino acids and, at the terminus, six charged amino acids. By enzymatically removing the parts of the M protein exposed above the cell wall, we have demonstrated that the section buried in the cell extends from about the last repeat of the C region to the C-terminus.

Reports from other laboratories have subsequently shown that similar arrangements of repeat blocks occur in the M proteins from type 5, 12 and 24 streptococci. An alignment of the amino acid sequences of these different M proteins reveals that their C-terminal ends are more than 98 percent identical. Closer to the N-terminus, however, differences in sequence among M proteins increase. Consequently, the A-repeat blocks and a short 11-amino acid region at the N-terminus are unique for each M molecule. Currently it seems that blocks of repeated sequences constitute many surface proteins found in gram-positive bacteria.

As my co-worker Belur N. Manjula and I continued to inspect the amino acid sequence of M6, another intriguing structural detail revealed itself. Running throughout all the repeat regions is an unusual seven-amino acid pat-



ROPELIKE STRUCTURE OF M PROTEINS (*left at top*) extends away from the surface of the streptococcus. Pairs of helical M molecules curl around one another, giving the protein its coiled-coil conformation. Specialized regions of the protein anchor it in the cell membrane and stabilize it in the cell wall. The cell wall (*above*) is a meshlike structure consisting of long chains of a disaccharide sugar that are randomly cross-linked by peptides.

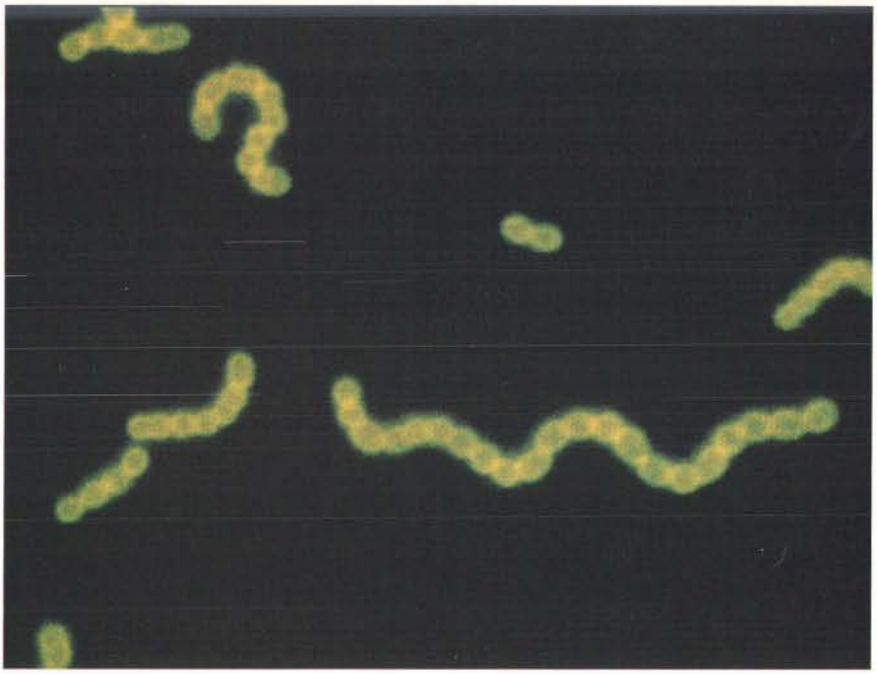
tern: the amino acids in the first and fourth positions are hydrophobic; the intervening amino acids allow the protein to twist itself into a spiral shape called an alpha helix. This observation prompted us to look through the literature for other proteins that shared this characteristic. All such molecules, we found, had an alpha-helical coiled-coil conformation—that is, they consisted of two alpha-helical chains twisted into a ropelike structure.

The coiled-coil structure forms because of the geometry of the alpha helix and the characteristics of its constituent amino acids. Making a single turn in an alpha helix requires 3.6 amino acids. Consequently, in the M6 protein the hydrophobic amino acids in the first and fourth positions line up to form an inclined stripe around the helix. Because of their hydrophobic nature, these amino acids prefer to be buried within a protein, away from the water molecules in the environment. Pairs of M6 alpha helices will therefore coil around one another to internalize, or cover, one another's hydrophobic amino acids.

The seven-unit pattern in the arrangement of the amino acids in M6 indicated that the repeat regions of the protein molecule make up a long helical rod. The pattern in M6 is not perfect, nor is that found in many other coiled-coil structures. Such irregularities probably account for the flexibility of the M molecules observed in electron micrographs. More important, the characteristics of these irregularities differ in the A-, B- and C-repeat regions. That observation suggests that each repeat region evolved independently and may have a distinct function. New data from a variety of ongoing experiments continue to verify that idea.

In collaboration with Carolyn Cohen of Brandeis University and George N. Phillips, Jr., who is now at Rice University, we gathered more physical and chemical data about M6. Our measurements confirmed that each M protein fiber on a streptococcal cell wall is about 50 to 60 billionths of a meter long and consists of a single coiled-coil molecule.

It is likely that M proteins of all serotypes are built along a basic theme: they have a lengthy coiled-coil rod region in their centers that is flanked by a floppy section at the N-terminal end and an anchoring region at the C-terminal end. Because the alpha-helical coiled-coil structure can accommodate a large number of varying amino acid sequences, many different M proteins with the same general conformation can be constructed. Natural selection pressures on the molecule appear to favor the coiled coil. Streptococcal mutants that produced M proteins with deviant structures would probably not survive in the human body.



STREPTOCOCCAL BACTERIA (*Streptococcus pyogenes*) grow in chains and can cause rheumatic fever. In this micrograph the streptococci have been labeled with a fluorescent dye by monoclonal antibodies that bind to the M protein.

For an M protein to protect a streptococcus, it must be able to attach to the organism. The mechanism that holds surface proteins on gram-positive bacteria is still poorly understood, but our studies of the M protein have been enlightening.

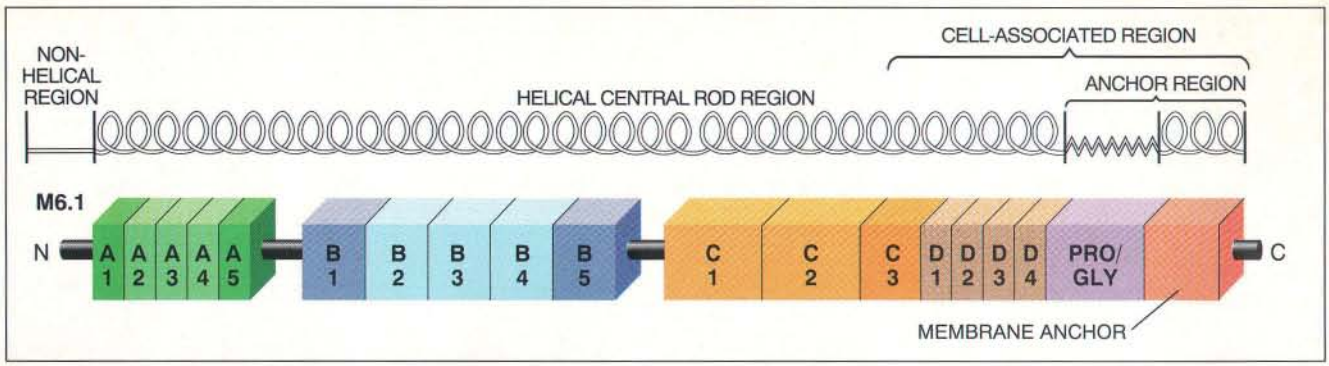
My Rockefeller colleague Vijaykumar Pancholi and I found that we could strip a streptococcus of its cell wall by exposing it to an enzyme that cuts the peptidoglycan. If we did so in a 30 percent solution of the sugar raffinose, the high osmotic pressure of the solution prevented the bacterial membrane from rupturing. Under those circumstances, the M protein remained attached to the exposed cell membrane. This simple experiment proved to us that the M protein is bound to the cell membrane and not linked to the cell wall, as many researchers had previously suspected.

We believed that the 20 hydrophobic amino acids near the C-terminal end were inserted into the similarly hydrophobic membrane itself, whereas the charged amino acids at the very terminus protruded into the aqueous cytoplasm. Because the charged amino acids would resist moving into a hydrophobic environment, they would act like a knot at the end of a string, preventing the M molecule from being pulled through the membrane.

That mechanism may be valid for some proteins attached to membranes. More recent evidence indicates, however, that the attachment mechanism for M protein and other bacterial surface proteins may actually be more sophisticated. Our studies have revealed that all surface proteins from gram-positive bacteria have a similar arrangement of hydrophobic and charged amino acids at their C-terminal end.

More important, however, a short six-amino acid sequence adjacent to the hydrophobic region is highly conserved in all the known surface proteins of gram-positive bacteria. The sequence consists of a leucine, a proline, a serine, a threonine, a glycine and a glutamic acid. Its designation is usually abbreviated as LPSTGE.

The importance of the LPSTGE sequence in the attachment of the M protein (and probably in all other proteins with this sequence motif) was dramatized by genetic experiments performed by Olaf Schneewind in my laboratory. He found that if he removed only the LPSTGE sequence from the M protein gene, the M molecule that was produced would not attach to the bacterial membrane. This result suggested that the hydrophobic domain and the charged amino acids at the C-terminus are not sufficient for membrane attachment and that the



PROTEIN SEQUENCE of the M6 molecule was determined by cloning its gene. About 80 percent of M protein consists of blocks of repeated amino acid sequences (A, B, C and D). The region rich in prolines and glycines may help the protein traverse the cell wall. Amino acids near the C-terminus permit attachment to the cell membrane.

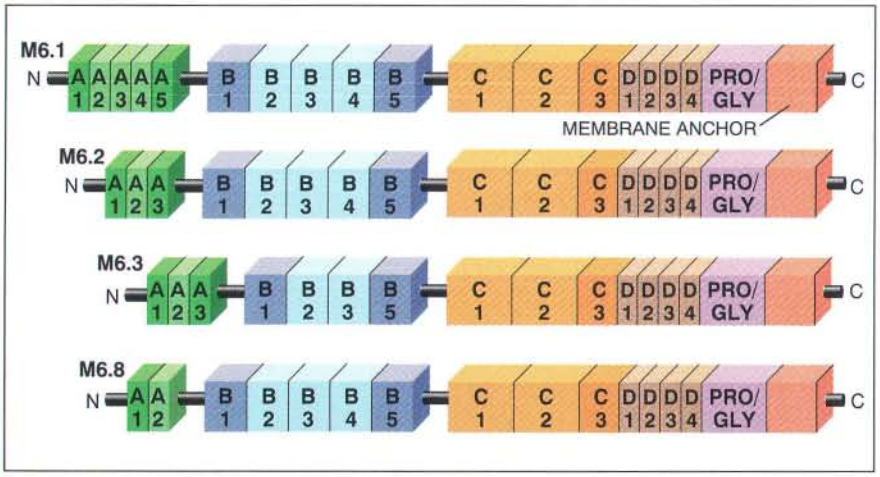
LPSTGE motif may be an important signal for initiating the process. The nature of the signal and the mechanism of attachment are currently under intense investigation.

In nearly all surface proteins found in gram-positive bacteria, there is another distinctive region that spans about 50 to 75 amino acids on the N-terminal side of the hydrophobic region. This part is probably located within the peptidoglycan. Proline, glycine, threonine and serine constitute a high percentage of these amino acids. The reason for their prevalence has not been fully explored, but it is known that prolines and glycines can create turns and bends in proteins. One hypothesis holds that cross-links in the peptidoglycan can weave through the proline- and glycine-induced bends, thereby stabilizing the M protein's position in the cell wall.

The discovery that all known surface proteins on gram-positive bacteria attach themselves by a similar mechanism may open new avenues for controlling infections caused by these organisms. Surface proteins help pathogenic organisms initiate infections. By preventing the proteins from anchoring to the bacterial cell, we should eventually be able to block infections and circumvent some of the problems associated with resistance to antibiotic therapies.

Just as the structures at the C-terminal end of the molecule tell us about how the M protein attaches to the bacterial cell, structures at the N-terminal end offer clues about how the molecule helps to fend off phagocytes. The N-terminal end of all M molecules has an excess of negatively charged amino acids, which results in a net negative charge for the region. Mammalian cells also exhibit a net negative charge on their surface. The charge on M proteins may thus have evolved to hamper contact between streptococci and phagocytic cells through electrostatic repulsion. It seems likely that one function of the central rod in the M protein is to act as a shaft for holding the negatively charged N-terminal end—and phagocytes—away from the bacterial surface.

At the N-terminal end of the coiled-coil rod, there is also a hypervariable region. This part of the molecule has a distinctive sequence in each M serotype. The hypervariable region consists of the short 11-amino acid nonhelical sequence and the adjoining A-repeat region. We found that the hypervariable region plays an important role in the biological activity of the molecule: only antibodies against this area permit phagocytes to consume the streptococci. That observation explains why only serotype-specific antibodies protect against strep infections.



DIFFERENT FORMS OF RELATED M PROTEINS arise when mutant streptococci delete copies of the amino acid repeats found in the parental molecules. Antigenic variation that results from these changes helps streptococci evade the immune system. Most variations in the M protein structures occur toward the N-terminus. The C-terminal half of the molecule is less variable.

For years, researchers have wondered how streptococci change the sequences in their hypervariable regions. Our finding that M molecules vary in size between organisms has led to a partial explanation.

Initially we examined 20 serotypes of M proteins that had been collected from streptococcal infections during a 40-year period. The molecules showed a wide range in size, from 41,000 to 80,000 molecular-weight units. We encountered a similar variation among streptococcal M6 proteins collected from several children in a daycare center during a pharyngitis outbreak.

To determine how rapidly the molecules change size, we examined the M proteins from samples of one M6 streptococcal strain that had been grown in our laboratory for more than 30 years. To our surprise, we did not find any evidence that its M protein had changed size during that time.

We also analyzed other streptococcal strains that had been serially collected from patients during a period of several weeks in the 1940s, before the advent

of penicillin therapy. Streptococci that had been isolated a week apart from the same patient had M proteins that were of the same serotype but were different in size. We therefore believe that in a patient's throat, natural selection pressures on the streptococci may favor the appearance of mutants that produce M proteins of different size. Such pressures would not exist for the strain growing in the laboratory.

Our further investigations revealed that size mutations are present in low numbers among the seemingly homogeneous cells of the M6 laboratory strain. According to our calculations, mutants that make smaller M proteins arise once in every 2,000 streptococcal chains. That rate is much too high to be explained by spontaneous point mutations, or changes in individual DNA bases. Such changes would be expected only about once in every one million to 10 million chains.

To examine what caused this change in size, we analyzed the M protein genes from several different mutants. Our results demonstrated that the changes had been caused by the removal of DNA sequences for certain repeat blocks in the M protein. For example, one mutant had deleted two identical A-repeat blocks from its M protein, and another mutant had eliminated one B-repeat block. In a third case the deletion extended from the center of the first A-repeat block to the center of the third A-repeat block. Because the end repeat blocks are slightly different, the deletion resulted in a change in the amino acid sequence.

These observations clarified how the changes in the M protein arose. During DNA replication in the multiplying streptococci, mistakes sometimes remove the DNA segments for some repeat blocks. As a result, the mutant offspring produce smaller M molecules. Larger M6 molecules appear in nature, but our laboratory techniques do not allow us to isolate them.

With this discovery in mind, my colleague Kevin Jones and I next asked whether antibodies that react with a parental M protein would also react with the shorter, mutant M molecules. Antibodies specific for the A and B repeat regions of a parental M6 protein provided us with a means to test the idea. We found that the antibodies either did not bind to the mutant M proteins or bound only very weakly if their normal binding sites were in a deleted area or immediately adjacent to one.

Size changes in M proteins can therefore interfere with the ability of some antibodies to bind to the molecule. Streptococci take advantage of genetic mistakes to change the size and antigenic character of their M molecules, thereby escaping recognition and destruction by the host's immune system. The streptococcal mutants that appeared in the weekly throat cultures of the pharyngitis patients may have survived because their M proteins were not targeted by the antibodies against the parental organisms.

This finding illustrates the basic strategy used by disease organisms to survive in the environment: they exploit their ability to replicate rapidly. One organism dividing every 30 minutes will produce more than a million daughter cells within 10 hours. Mistakes during this replicative process create mutant organisms. Usually a mutant will not survive, because its defect is in a function vital for growth. In other instances, however, the genetic change may allow the mutant to survive and prosper in an environment that is unsuitable for other daughter cells.

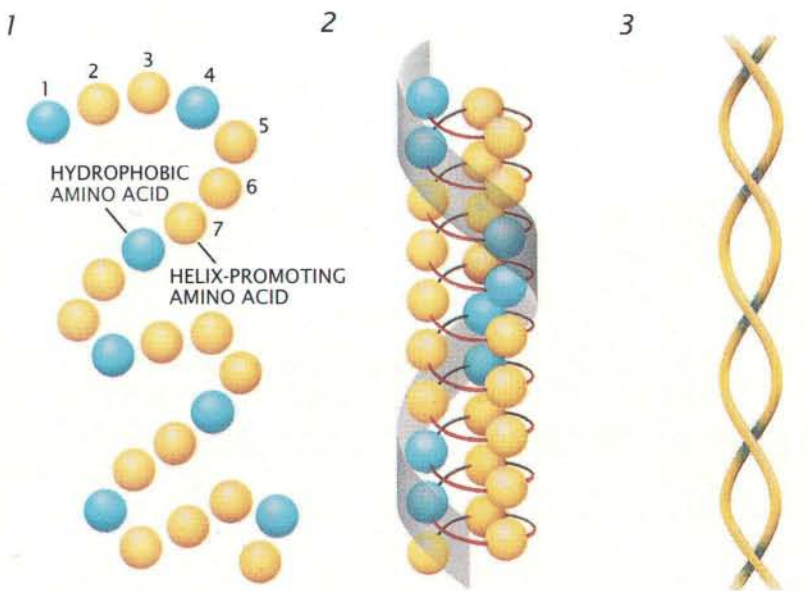
The same principle applies to the M protein and other surface molecules with the capacity to change. Often DNA changes occur in genetic "hot spots," where repeat sequences predominate. Because those areas are particularly prone to mistakes during DNA replication, the microorganism maintains the repeat sequences to ensure a steady supply of mutants. Although this mechanism explains how changes occur in the repeat regions, we still do not understand how the 11 amino acids at the N-terminus change between serotypes.

I have now outlined two ways in which the M protein may confound a human host's defenses: N-terminal repulsion and antigenic variation. To understand a third mechanism, we must first review another surveillance system by which the host identifies and prevents microbial infection.

In their blood and lymphatic fluid, mammals carry a complex array of proteins and enzymes—the complement system—that recognizes pathogenic microorganisms and other foreign particles. When a foreign particle enters the body, the complement protein C3b quickly binds to it. The C3b label marks the pathogen as something to be killed and cleared from the body by white blood cells or other related systems. Such a rapid process for recognizing invaders can make mistakes by wrongly labeling normal mammalian cells. Therefore, the blood serum also contains a set of regulatory proteins that can reverse or block the binding of C3b to these cells.

While investigating why streptococci are not cleared after an infection, researchers noticed that C3b bound effectively to streptococci without M proteins but not to M-bearing ones. In collaboration with Rolf Horstmann of the Bernhard-Nocht Institute in Hamburg,

How M Protein Becomes a Coiled Coil



The M protein sequence has a regular pattern of hydrophobic amino acids.

When the sequence coils into an alpha helix, a hydrophobic stripe forms.

Pairs of helical strands intertwine to cover their hydrophobic regions.

we discovered the reason for this difference: factor H, one of the regulatory proteins of the complement system, can bind specifically to the M protein. That binding limited the deposition of C3b onto the streptococci and protected them from being cleared. In a sense, the M-bearing streptococcus cleverly disguises itself as a normal human cell to evade the complement system. We found subsequently that factor H binds to the M protein within the conserved C-repeat region.

That strategy for evading clearance from the body has made the streptococcus a successful human parasite. We postulate that during an infection, factor H binds to the M protein, thereby preventing the destructive labeling of the streptococci. If the infection goes untreated, the host will generate antibodies against the M protein within seven to 14 days.

We believe that antibodies that bind to the B- and C-repeat regions of the molecule fall under the influence of factor H. Consequently, those antibodies do not help to eliminate the infection. Only antibodies against the hypervariable region near the N-terminal end are far enough away from factor H to trigger the destruction of the bacteria by antibody-mediated phagocytosis. Antibodies to the N-terminus also neutralize the region's negative charge and thereby assist phagocytosis.

The full cycle of infection, antibody production and clearance from the body requires up to 14 days. During that time the streptococcus can generate a large number of progeny, some of which will be identical to their parent, but some of which will be mutants that have changed M proteins. The successful cells are usually passed on to another individual to begin a new cycle and to perpetuate the species.

Our structural studies have helped point out how a group A streptococcal infection may indirectly cause the heart damage associated with rheumatic fever. We compared the amino acid sequences of M proteins with those of other proteins that were recorded in a computer data base. The M proteins were as much as 40 percent identical with other fibrous coiled-coil proteins such as tropomyosin, myosin and keratin—all three of which are found in mammalian tissues. That similarity has immunologic consequences.

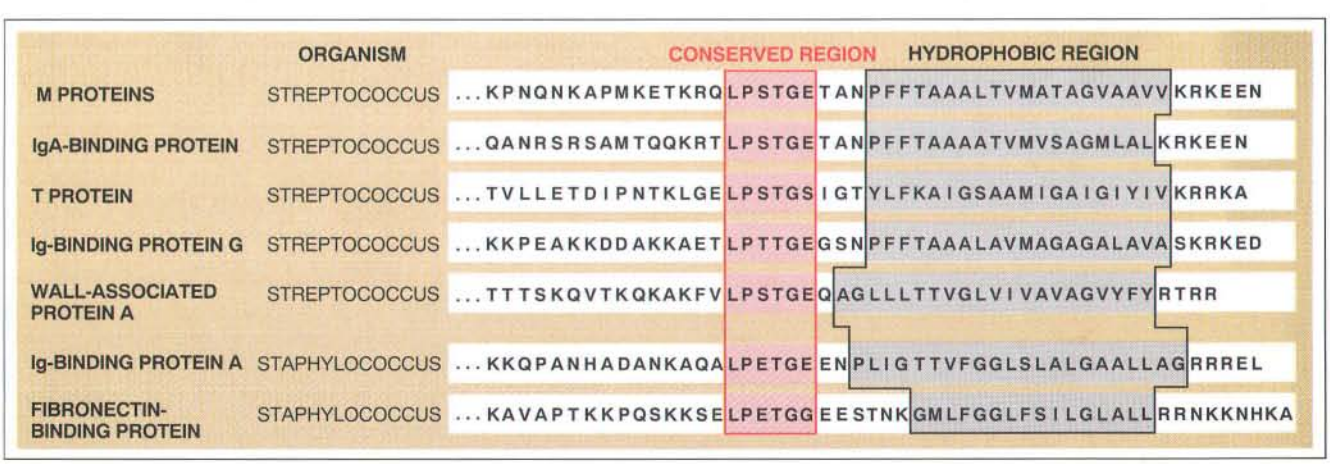
One hallmark of rheumatic fever is the presence of antibodies that react with muscle tissue, particularly heart tissue, in a patient's blood serum [see "Rheumatic Fever," by Earl H. Freimer and Maclyn McCarty; SCIENTIFIC AMERICAN, December 1965]. Normally, antibodies are not made against one's own tissues. Researchers have discovered, however, that so-called cross-reacting antibodies can sometimes be induced by a molecule in an infective organism that resembles one in the mammalian host. In the process of making antibodies against the microbial molecules to clear an infection, the body is tricked into generating antibodies against its own tissues—a potentially harmful development.

The late Edwin H. Beachey and his co-workers at the University of Tennessee at Memphis found that rabbits immunized with purified M protein produced antibodies to both the M protein and the mammalian muscle myosin. Some antibodies produced in this way have also been found to react with tropomyosin and other mammalian muscle proteins. Because these muscle proteins are also coiled-coil molecules, it is likely that the cross-reactive antibodies recognize a common conformational feature of the M protein and the mammalian molecules. The exact role of such cross-reactive antibodies in the genesis of rheumatic heart disease, however, is not yet understood.

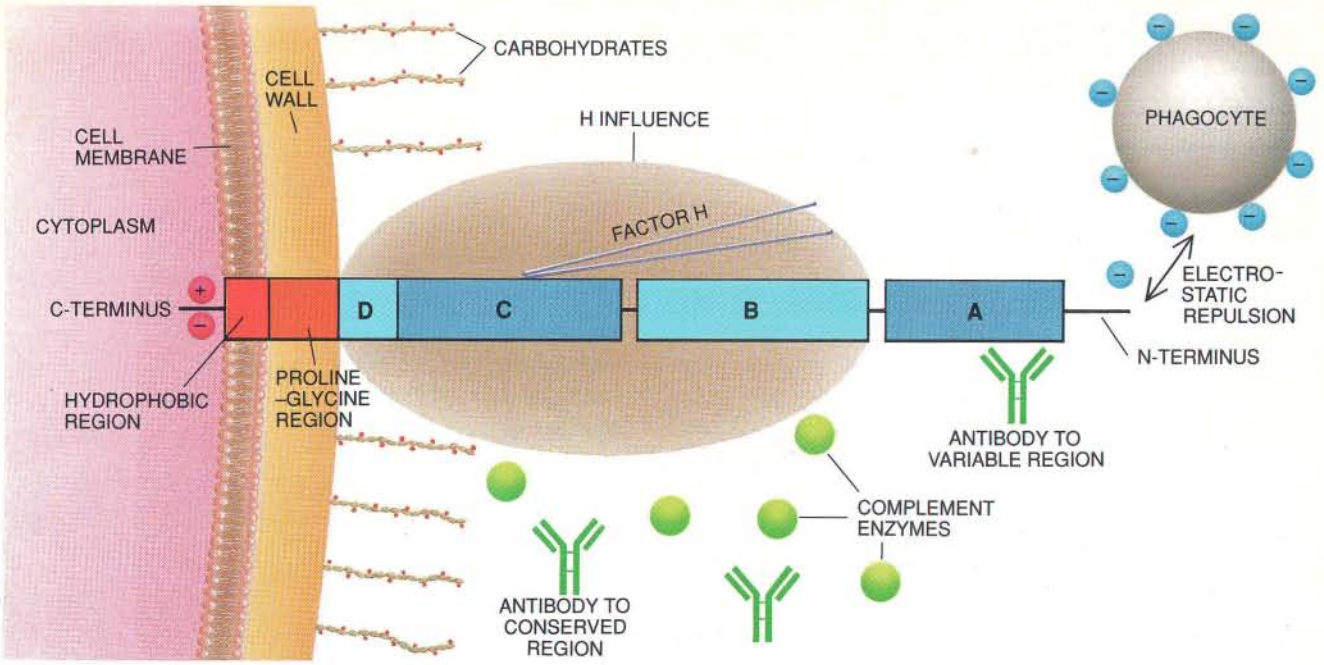
Attempts have been made to apply the knowledge of the M protein's structure and function to the development of a vaccine for preventing streptococcal infections. Antibodies against the N-terminal hypervariable region will initiate phagocytosis, but as mentioned previously, vaccines prepared from that region will protect against only a single type of streptococcus. Such vaccines would therefore need to include many different N-terminal sequences. That problem, along with the continuously changing nature of the N-terminal region, makes the approach unattractive.

An alternative strategy that we have explored takes advantage of the conserved regions of the M molecule that are not buried within the cell wall or membrane. This tactic was inspired by the observation that the incidence of group A streptococcal pharyngitis generally peaks at about age six or seven, then declines rapidly after age 10. The lowest incidence of pharyngitis occurs after age 20. Yet it is unlikely that adults have been exposed to all of the more than 80 streptococcal types, and recent experiments in my laboratory have verified that suspicion.

Rather, the lower incidence of streptococcal pharyngitis during adulthood may be caused either by an undefined age-related factor or by a broad immunity that individuals acquire through contact with streptococci as children. We postulated that protection against pharyngitis could be induced by anti-



SIMILAR REGIONS appear near the C-termini of several proteins from various gram-positive bacteria. Each letter represents an amino acid in the protein sequence. The similarities at the C-termini suggest that a common mechanism is responsible for attaching the C-ends of M proteins and of these other molecules to bacteria.



THWARTING THE IMMUNE SYSTEM is the primary job of the M protein. Negative charges at the N-terminus may repel phagocytic white blood cells. By binding with factor H—a regulatory protein produced by the human host—the M protein protects its most conserved regions from antibodies and complement enzymes. Only antibodies against the antigenically shifting hypervariable region can clear an established streptococcal infection from the host's body.

bodies to some regions that are conserved among several serotypes. In this way, exposure to streptococci during childhood may permit the development of a repertoire of antibodies against the conserved regions of M proteins, and these antibodies confer protection later in life.

To test this hypothesis, my colleague Debra Bessen and I prepared synthetic peptides of about 20 amino acids. Each peptide was a copy of an amino acid sequence found in the conserved C-repeat region of the M protein. These peptides were then coupled to a nontoxic subunit of the cholera toxin molecule. The subunit served as a carrier for the small peptides and helped to stimulate an immune response against them. This protein-peptide complex was sprayed into the noses of mice several times during a two-week period and again after three weeks as a booster. We waited a short time to allow the mice to develop appropriate antibodies against the peptides, and then we challenged them with intranasal and oral exposures to live group A streptococci.

The animals that had received the peptide vaccine, we found, were significantly less prone to streptococcal pharyngitis than mice that had not. In essence, the vaccine prevented the mice from getting strep throat. The antibodies against the conserved regions of the M protein apparently blocked the streptococci from attaching to the pharynx (throat tissue) and colonizing it. We saw the protective effect even when the challenging organism had an M serotype different from that in the vaccine.

To test a different approach, in collaboration with Dennis E. Hruby of Oregon State University, we transferred the gene for the conserved region of the M protein into a vaccinia virus. (These viruses, which are the basis for the vaccine against smallpox, are weak and relatively harmless to humans.) When the genetically altered viruses infected mammalian cells, they caused the cells to produce the conserved region of the M protein. We vaccinated mice intranasally with the viruses. Thereafter the mice were immune to challenges by live streptococci.

In summary, it seems possible to sensitize animals' immune systems against the conserved regions of M proteins. By preventing the initial colonizing step of an infection, we can circumvent the need for type-specific antibodies. Type-specific antibodies become necessary for overcoming the infection only after streptococci have colonized and invaded tissues.

Since Lancefield first identified the M protein as an important determinant of the virulence of streptococci, it has taken scientists almost 60 years to develop some form of protection against the infections. Several more years will probably pass before a vaccine for humans is ready. The progress has hinged on the growing understanding of the molecular structure of the M protein, the location of its functional domains and the method by which the protein repels phagocytic attack. It is exciting to realize that the studies of the M protein will benefit not only those persons at risk for streptococcal infections and rheumatic fever: they should also prove useful in developing strategies against other bacterial and viral diseases.

FURTHER READING

STREPTOCOCCAL M PROTEIN: ALPHA-HELICAL COILED-COIL STRUCTURE AND ARRANGEMENT ON THE CELL SURFACE. George N. Phillips, Jr., Paula F. Flicker, Carolyn Cohen, Belur N. Manjula and Vincent A. Fischetti in *Proceedings of the National Academy of Science*, Vol. 78, No. 8, pages 4689–4693; August 1981.

EPITOPES OF STREPTOCOCCAL M PROTEINS SHARED WITH CARDIAC MYOSIN. James B. Dale and Edwin H. Beachey in *Journal of Experimental Medicine*, Vol. 162, No. 2, pages 583–591; August 1, 1985.

STREPTOCOCCAL M PROTEIN SIZE MUTANTS OCCUR AT HIGH FREQUENCY WITHIN A SINGLE STRAIN. Vincent A. Fischetti, Mary Jarymowycz, Kevin F. Jones and June R. Scott in *Journal of Experimental Medicine*, Vol. 164, No. 4, pages 971–980; October 1, 1986.

PROTECTION AGAINST STREPTOCOCCAL PHARYNGEAL COLONIZATION WITH A VACCINIA: M PROTEIN RECOMBINANT. Vincent A. Fischetti, Walter M. Hodges and Dennis E. Hruby in *Science*, Vol. 244, pages 1487–1490; June 23, 1989.

Polar Stratospheric Clouds and Ozone Depletion

Clouds rarely form in the dry, Antarctic stratosphere, but when they do, they chemically conspire with chlorofluorocarbons to create the "ozone hole" that opens up every spring

by Owen B. Toon and Richard P. Turco

More than two dozen scientists boarded a National Aeronautics and Space Administration's DC-8 based at Punta Arenas, Chile. The aircraft headed south, climbing through the sunrise sky high above the Antarctic peninsula into the stratosphere, that part of the atmosphere between 10 and 50 kilometers in altitude. As the plane entered the now well-known ozone hole in that September of 1987, it was greeted by a large cloud in the shape of an eye, with a bright red iris surrounding a green pupil. Along with an ER-2, a companion high-altitude aircraft, the DC-8 carried instruments to measure the aerosols, gases and atmospheric dynamics in such clouds as well as in the surrounding stratosphere. The expedition would help scientists to understand what had been observed two years before: the correlation between the depletion of ozone and the formation of those curious clouds.

During the past century, observers on land had periodically recorded the appearance of stratospheric clouds over both poles, at an altitude of about 20 kilometers. The clouds extend 10 to 100 kilometers in length and several kilometers in thickness. They glow with a seashell iridescence. Hence they are sometimes called nacreous, or mother-of-pearl, clouds.

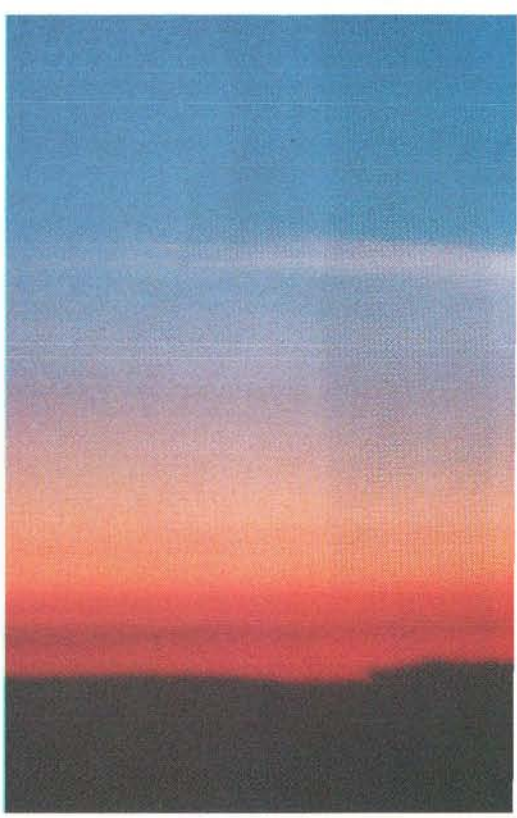
In addition to nacreous clouds, investigators have found two other types. The second kind of cloud consists of nitric acid instead of pure water. The third species is identical to nacreous clouds in chemical composition but forms in a process that results in a larger cloud with no iridescence. When these three kinds of clouds form over the poles, scientists broadly refer to them as polar stratospheric clouds—PSCs, for short.

The eerie beauty and exotic nature of PSCs belie their more ominous significance. Recent work, including our own research, strongly indicates that PSCs trigger ozone depletion in the Arctic stratosphere. In the Antarctic stratosphere they help to create the ozone hole.

The ozone hole is not actually a hole but a region that contains an unusually low concentration of ozone [see "The Antarctic Ozone Hole," by Richard S. Stolarski; SCIENTIFIC AMERICAN, January 1988]. At ground level this molecule, consisting of three oxygen atoms, is a health hazard. In the stratosphere ozone is crucial to survival on the earth. Although ozone constitutes less than one part per million of the gases in the atmosphere, it absorbs most of the ultraviolet radiation from the sun.

Such radiation can affect the growth and reproduction of phytoplankton, the base of the marine food chain. In humans, excessive ultraviolet exposure has been implicated as a cause in skin cancers, cataracts and immune deficiencies. Although remotely located, the ozone hole in the Antarctic is nonetheless cause for concern: shifting circulation patterns carry masses of ozone-depleted air north. It may thus forebode widespread ozone depletion throughout the stratosphere.

In 1985 Joseph C. Farman and his coworkers at the British Antarctic Survey first reported that significant ozone depletion had been occurring over Antarctica since the late 1970s. Satellite measurements from *Nimbus 7*, managed by Arlin Krueger of the NASA Goddard Space Flight Center, showed that over the years the depletion from austral spring to austral spring has generally worsened. About 70 percent



STRATOSPHERIC CLOUDS can form over the polar regions if the air cools suffi-

OWEN B. TOON and RICHARD P. TURCO study atmospheric chemistry, radiation and microphysics, frequently applying information gained from planetary explorations. Toon received his Ph.D. from Cornell University and is currently an associate fellow at the National Aeronautics and Space Administration Ames Research Center. He was one of the leaders of the airborne missions to the polar regions. Turco received his Ph.D. from the University of Illinois at Urbana-Champaign and is professor of atmospheric sciences at the University of California, Los Angeles. The authors were members of the research team that first introduced the concept of nuclear winter.

of the ozone above Antarctica, which equals about 3 percent of the earth's ozone, is lost during September and October. Measurements by David Hofmann and his co-workers at the University of Wyoming revealed that most of the ozone loss occurred at altitudes between about 12 and 30 kilometers.

Scientists have advanced a number of theories to explain what causes the ozone hole. Several major expeditions have served to winnow out the wrong ones. One principal group of theories, for example, suggested that atmospheric motions alone caused the ozone hole. Proponents of these theories thought that the circulation pattern over the poles may have gradually changed, so that upward moving winds might now blow over Antarctica during the spring. These winds would replace ozone-rich stratospheric air with ozone-poor air from the troposphere, the atmosphere below 10 kilometers.

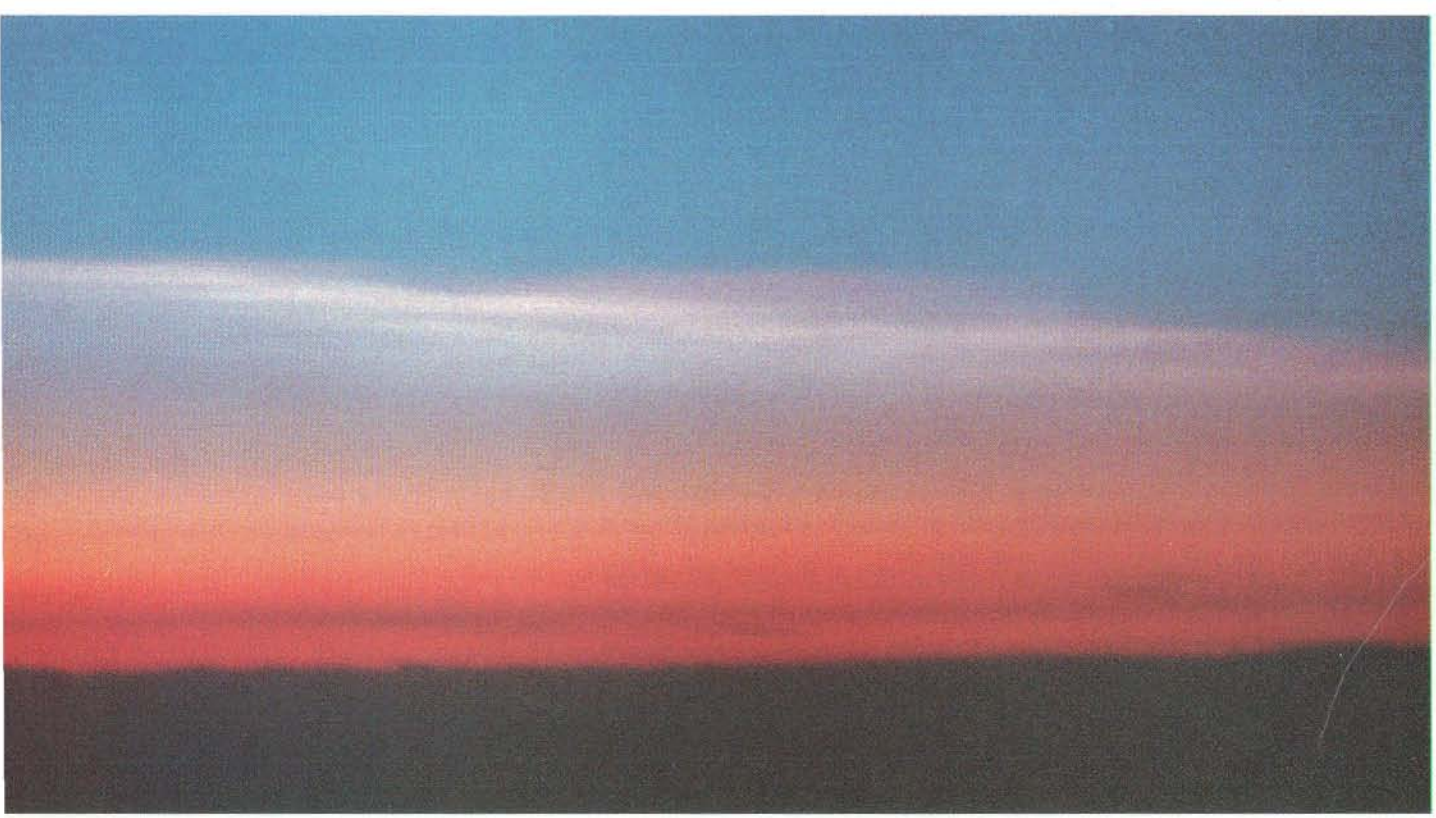
Max Loewenstein and his group from the NASA Ames Research Center, Leroy E. Heidt and his colleagues at the National Center for Atmospheric Research (NCAR) and others showed such hypotheses to be incorrect. According to the dynamic models used by advocates of the circulation theories, high concentrations of trace gases originating from the ground should be present at the altitude of the ozone hole. Measurements by the investigators revealed only low levels of the trace gases, however, indicating that ozone-hole air in fact comes from high altitudes, where ozone is normally abundant.

A second class of theories proposed that chemical reactions deplete ozone. One early hypothesis suggested that reactive nitrogen compounds, normally the most important agents for destroying ozone in the lower stratosphere, might exist at elevated concentrations near the ozone hole. The enhancement was presumed to result from the combined effects of increased solar activity and atmospheric circulation.

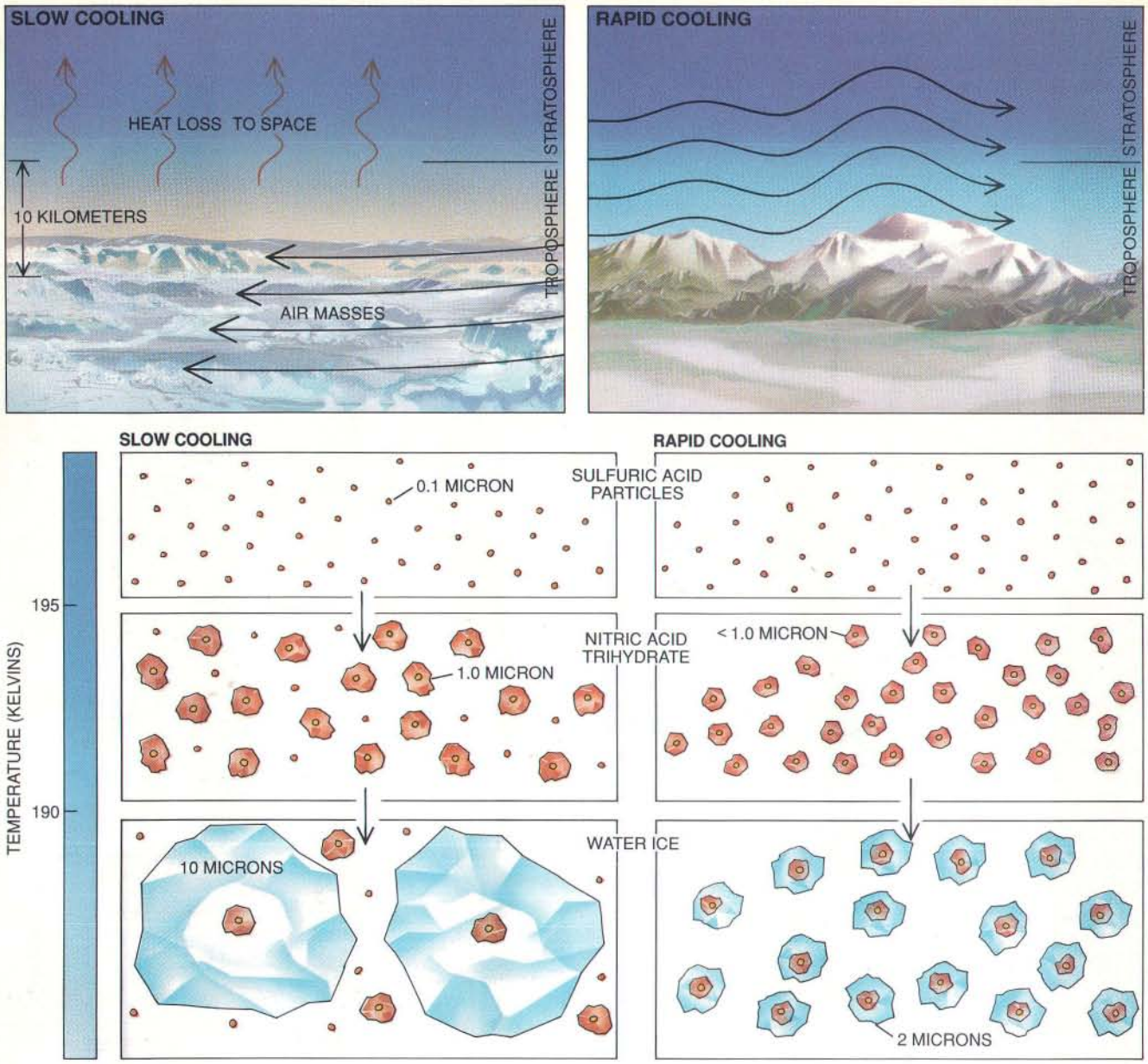
The theory proposed that the enhanced solar activity produced reactive forms of nitrogen over the South Pole at high altitudes. The downward motion of air then carried the reactive nitrogen into the lower stratosphere, where investigators observed the ozone loss. But Crofton B. Farmer and his colleagues from the NASA Jet Propulsion Laboratory, George H. Mount and his co-workers at the National Oceanic and Atmospheric Administration (NOAA) Aeronomy Laboratory and others found that the reactive forms of nitrogen were also depleted in the ozone hole, hence disproving the theory.

Farman and his colleagues proposed an alternative chemical interpretation, one that has now gained wide acceptance. Based on the mid-1970s work by Mario J. Molina, now at the Massachusetts Institute of Technology, and F. Sherwood Rowland of the University of California at Irvine, the theory suggests that chlorine compounds might be responsible for the ozone hole. Chlorine primarily enters the atmosphere as a component of chlorofluorocarbons (CFCs) produced by humans. These inert compounds, used in such diverse applications as coolants for air conditioners and refrigerators, as solvents for cleaning circuit boards and as agents for producing insulating foams, may survive 50 to 100 years in the atmosphere. In only a few years, winds throughout the troposphere uniformly distribute CFC molecules released from a single point. Over the decades, the molecules eventually reach the middle stratosphere, about 30 kilometers or higher. There, ultraviolet light from the sun tears them apart.

The chlorine released from the CFC molecules initially either exists as free chlorine or reacts with ozone to form chlorine monoxide (ClO). These two forms of chlorine then react further to form stable compounds that are the so-called reservoirs of chlorine. The reser-



ciently during the winter. Nitric acid trihydrate clouds are visible as thin, dark orange layers. The water-ice clouds appear whitish. The clouds, such as these over Stavanger, Norway, help to initiate the chemical reactions that destroy ozone.



FORMATION OF POLAR STRATOSPHERIC CLOUDS occurs when the air cools sufficiently for vapor to condense. The cooling may take place slowly, as when the stratosphere radiates heat into space or is uplifted by air masses sliding below it (*top left*). Rapid cooling occurs when air flows over mountains, creating a standing-wave pattern that reaches the stratosphere (*top right*). As the temperatures drop below 195 kelvins, particles of nitric acid trihydrate condense around the sulfuric acid particles in the stratosphere, forming nitric acid trihydrate clouds. Water-ice clouds form if the temperature drops below 190 kelvins; the water vapor condenses on the nitric acid trihydrate particles and on any remaining sulfuric acid particles. Under rapidly cooling conditions, a greater fraction of such particles become condensation nuclei.

voirs consist primarily of the gaseous forms of hydrochloric acid (HCl), produced in a reaction of free chlorine with such common atmospheric constituents as methane, and chlorine nitrate ($ClONO_2$), formed in a reaction between ClO and nitrogen dioxide (NO_2).

Chlorine reservoirs themselves do not destroy the ozone layer. In these compounds, chlorine remains inert and cannot react with ozone. Early computer models concluded that CFCs should not have a major effect on the ozone layer. They indicated that only small amounts of ozone would be destroyed by some of the chlorine in the reservoirs that does manage to escape and become active.

Evidently some mechanism in the Antarctic stratosphere was freeing more of the chlorine from these inert reservoirs. Susan Solomon and her co-workers at the NOAA Aeronomy Laboratory and Michael B. McElroy and his coworkers at Harvard University provided the first hints of what this mechanism might be. In 1986 they suggested that the observed correlation between the cycle of ozone depletion and the presence of polar stratospheric clouds implied that chemical reactions taking place on the ice particles in the clouds freed chlorine from the reservoirs.

At first glance, the cloud theory encountered an apparent problem: clouds in the stratosphere were thought to be uncommon. The relative humidity there averages about 1 percent. Moreover, water vapor constitutes only a few parts per million of the air, a factor of 1,000 less than the

amount in the troposphere, where most clouds form.

Until recently, the only kind of stratospheric cloud thought to exist was the nacreous cloud. These clouds form at altitudes of about 15 to 30 kilometers and are the stratospheric versions of the lenticular (lens-shaped) clouds familiar to inhabitants of windy, mountainous regions. Lenticular clouds form as air rushes over mountains. The air creates a standing pattern of so-called lee waves downwind from the mountains. In the ascending portion of the lee waves, the air rapidly expands and cools. If there is sufficient moisture, it will condense on the many particles in the air. The waves thus become visible as clouds.

If the air is stably stratified and the wind does not change speed or direction at higher altitudes, the standing-wave pattern created by the mountains can propagate into the stratosphere. Nacreous clouds then tend to form on the crests of the standing waves. They do so through condensation on any aerosols present.

Sudden cooling and condensation of water vapor form the nacreous clouds. Because the mountains create standing waves, the clouds remain stationary, even though air constantly rushes through them. The ice crystals collect water as air currents push them through the clouds, so they grow to about two microns in size before all the water has been collected. When the crystals reach the descending portion of the lee waves, the air compresses and thus heats, evaporating the ice. A single cloud may extend from 10 to 100 kilometers in length. The energy of the standing wave may be sufficiently high so that a succession of such clouds may exist.

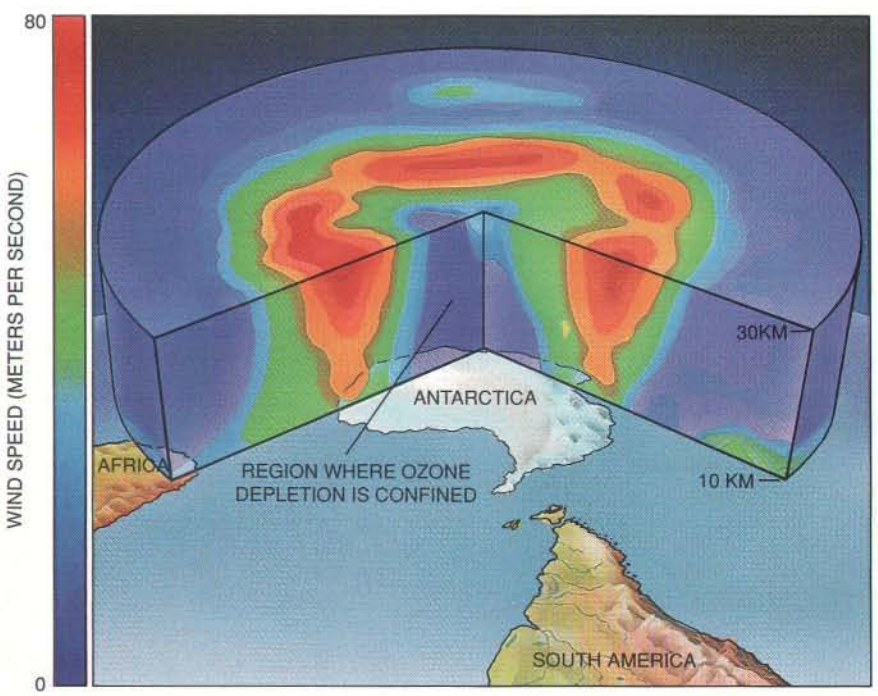
The distribution of sizes across the cloud gives nacreous clouds their iridescence. The smallest ice particles occur at the leading and trailing edges because the particles there have just begun to grow or have nearly evaporated; the largest form at the center. Particles at all positions in the cloud diffract passing sunlight. The intensity of the diffracted light depends on the wavelength of the light and size of the particle. As a result, when the clouds are viewed at moderate angles from the direction of the sun, they appear brightly colored. The colors follow the contour of the clouds, mimicking the distribution of particle sizes.

Nacreous clouds indicated to meteorologists that the stratosphere was indeed cold enough to enable water ice to form, at least near the polar regions. But because of the extreme dryness, the temperature must fall below 190 kelvins (–83 degrees Celsius). Only during the Antarctic winter are such temperatures maintained for any length of time. (Clouds also exist in the Arctic stratosphere, but they form less frequently because the average winter temperatures there are higher than in the Antarctic.)

Satellite data, however, revealed what land-based observers could not see. The Stratospheric Aerosol Measurement (SAM) II instrument, launched on board the *Nimbus 7* satellite in 1978 and managed by M. Patrick McCormick of the NASA Langley Research Center, detected particles in the air by examining sunlight as it grazes the limb of the earth. SAM II showed that stratospheric clouds existed over Antarctica even when the temperature dropped to only 195 kelvins (–78 degrees C). Such temperatures are too warm for nacreous clouds to form. Therefore, the clouds can be supposed to be created by some other process. Buttressing that conclusion was the fact that these newly discovered PSCs were too extensive to have formed from air currents flowing over mountains.

Along with Paul J. Crutzen of the Max Planck Institute for Chemistry in Mainz and Frank Arnold of the Max Planck Institute for Nuclear Physics in Heidelberg, we and our co-workers proposed in 1986 that these clouds must differ in composition from nacreous clouds, which consist of pure water condensed on suspended particles. Chemical theories for ozone destruction require the removal of reactive nitrogen, which would otherwise trap chlorine as chlorine nitrate, one of the primary chlorine reservoirs. We deduced that the clouds might serve as a nitrogen sink. If so, they would consist of a frozen form of nitric acid (HNO_3) with three water molecules for each nitric acid molecule. Such a compound, called nitric acid trihydrate ($HNO_3 \cdot 3H_2O$), not only accounts for the nitrogen removal but also condenses at temperatures higher than does pure water.

In a series of independent observations, groups led by David W. Fahey of the NOAA Aeronomy Laboratory, Bruce W. Gandrud of NCAR and Rudolf F. Pueschel and Stefan A. Kinne of the NASA Ames Research Center confirmed our theories. Along with other researchers, we also determined that these clouds do not commonly form by the sudden cooling of air uplifted by mountains. A slowly cooling process usually produces these clouds. The winter polar stratosphere radiates energy away to space through the long polar night, and eventually vast regions reach temperatures at which cloud particles will form. In addition, weather systems in the lower atmosphere slide beneath



ANTARCTIC POLAR VORTEX confines the ozone depletion within a ring of rapidly circulating air. Wind speeds are represented by colors within the disk. The geography is not to scale. The diagram is based on a computer-generated image made by Mark R. Schoeberl and Leslie R. Lait of the National Aeronautics and Space Administration Goddard Space Flight Center.

- Antarctic winter begins.
- Vortex develops, and temperature becomes cold enough for clouds to form.
- PSCs denitrify and dehydrate the stratosphere through precipitation.
- Hydrochloric acid and chlorine nitrate react on cloud surfaces to free chlorine.
- Winter temperatures drop to their lowest point.

the polar stratosphere, lifting and cooling the air. During the winter, when these processes cause the temperature to drop below about 195 kelvins, clouds of nitric acid trihydrate form. The sulfuric acid particles in the air serve as seeds. Such seed particles come from sulfur gases produced by natural biological processes and anthropogenic sources. Circulation patterns transport the sulfur released in the lower atmosphere to the stratosphere. Also, explosive volcanic eruptions can spew sulfur gases directly into the stratosphere. The particles, about 0.1 micron in size, may be especially abundant for a few years thereafter. In 1982 El Chichón in Mexico, near Pichucalco, expelled about five million tons of sulfur into the stratosphere.

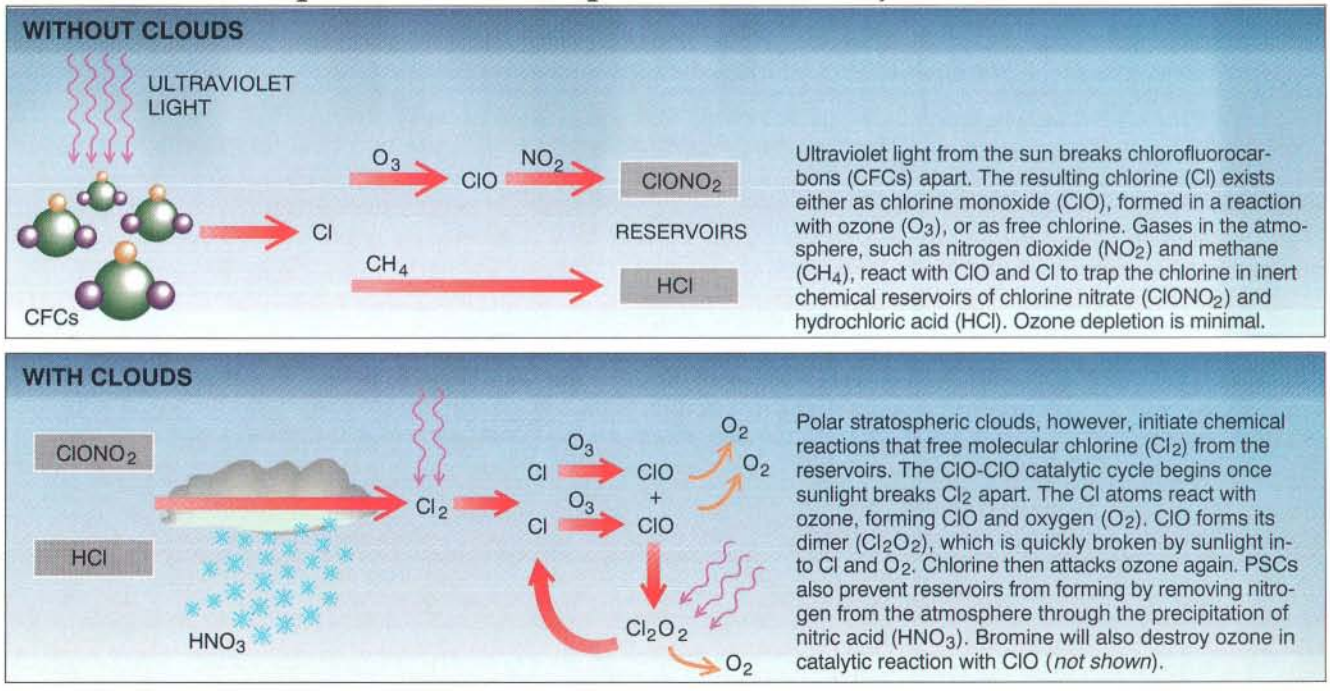
The slow cooling can produce geographically extensive stratospheric clouds. McCormick and Edward V. Browell of the NASA Langley Research Center and their co-workers used aircraft-borne laser radar (lidars) to map individual clouds. They found that the clouds often occur as multiple, kilometer-thick layers stretching out over distances sometimes exceeding several thousand kilometers. Compared with nacreous clouds, nitric acid trihydrate clouds are less massive and more tenuous, making them difficult to see with the naked eye.

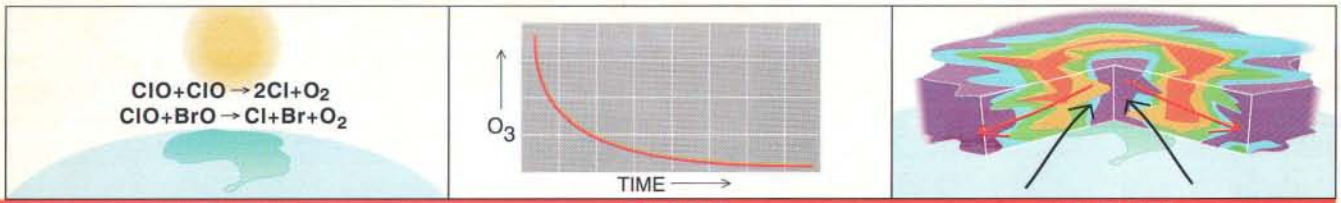
In addition to nacreous and nitric acid trihydrate clouds, another kind of PSC can form. The third type occurs if the Antarctic winter temperature slowly drops below 190 kelvins. As the air cools, water vapor condenses on some of the suspended particles, forming water-ice clouds. The seed particles are the nitric acid particles (which have themselves grown on sulfuric acid particles) that compose the nitric acid trihydrate clouds.

This type of PSC, like the nacreous clouds, contains water ice. Researchers commonly classify the two kinds of water-ice clouds together but distinguish them by their rate of formation (nacreous clouds form by rapid cooling). These water-ice clouds are not as common as nitric acid trihydrate clouds, especially in the Arctic, because of the extremely low temperature necessary for their formation.

Unlike the rapidly cooling water-ice clouds, the slowly cooling water-ice clouds are barely visible to ground observers. Although the mass of condensed water is nearly equal in both kinds of water-ice clouds, the particles in the slowly cooling clouds are larger than those of nacreous clouds. The rapid cooling that forms nacreous clouds transforms virtually all suspended aerosols into nuclei for condensation; slow cooling uses only a small fraction of the particles present. As a result, nacreous clouds contain a large number of small ice crystals, about two microns in size. Slowly cooling clouds have fewer, but larger, crystals that exceed a size of 10 microns. Because slowly cooling water-ice clouds contain fewer discrete particles per volume, they do not reflect light as well as their nacreous cousins do.

How Polar Stratospheric Clouds Help Chlorine Destroy Ozone





- Sunlight returns to the center of vortex as the austral spring begins, and PSCs disappear because of increasing temperatures.
- ClO-ClO and ClO-BrO catalytic cycles destroy ozone.

- Lowest levels of ozone are reached.

- Polar vortex breaks down.
- Ozone-rich air from mid-latitudes replenishes the Antarctic stratosphere.
- Ozone-poor air spreads over the Southern Hemisphere.

All three types of PSCs—nitric acid trihydrate, slowly cooling water-ice and rapidly cooling water-ice (nacreous) clouds—act as key components of the Antarctic ozone depletion. The PSCs can activate chlorine on their surfaces as well as use up reactive nitrogen, which would otherwise transfer chlorine to its reservoirs. Furthermore, slowly cooling water-ice and nitric acid trihydrate clouds can entirely deplete the stratosphere of nitrogen.

Laboratory studies by Molina and Ming-Taun Leu of J.P.L. and Margaret Tolbert of the Stanford Research Institute International and their co-workers showed that a reaction between hydrochloric acid and chlorine nitrate—the two compounds that hold chlorine in its inactive state—will indeed occur on water-ice and nitric acid trihydrate surfaces. That reaction, which produces molecular chlorine (Cl_2) and nitric acid, is negligibly slow without the presence of solid particles.

In the sunlight of the Antarctic spring, molecular chlorine quickly dissociates into highly reactive atomic chlorine, precipitating the ClO-ClO catalytic cycle. In this cycle the newly liberated chlorine atom breaks apart ozone to yield an oxygen molecule and chlorine monoxide. Molina discovered that the gas-phase chlorine monoxide reacts with itself, forming its dimer (Cl_2O_2). Sunlight readily dissociates the dimer into free chlorine atoms, leading to further ozone destruction. Chlorine thus maintains a catalytic role in ozone depletion [*see illustration on opposite page*].

Were nitrogen dioxide present, it would quickly combine with chlorine monoxide to trap the chlorine in the inert reservoir molecule, chlorine nitrate, thus halting the ClO-ClO catalytic cycle. But the PSCs prevent the reaction because they convert any nitrogen present into nitric acid.

James G. Anderson and his co-workers at Harvard University, as well as Robert L. deZafra and Philip Solomon of the State University of New York at Stony Brook, found astonishingly high levels of ClO in the Antarctic ozone hole—about 500 times the amount found at mid-latitudes at the same altitude. Given such a high concentration, the ClO-ClO catalytic cycle can account for most of the observed losses in the ozone hole. A single chlorine atom may destroy thousands of ozone molecules before encountering reactive nitrogen or hydrogen compounds that eventually return chlorine to its reservoirs. Laboratory studies of these various reactions are continuing so that scientists can refine the pathways the chemistry takes and determine more precisely the reaction rates.

The active chlorine freed by the PSCs also plays a role in another significant catalytic process, one that involves bromine. Human activities contribute bromine, which is an important component of some types of fire-extinguishing compounds, to the atmosphere. The reaction may account for about 20 percent of the ozone destruction.

McElroy and his co-workers first suggested how the reaction proceeds. Bromine removes an oxygen atom from ozone, forming bromine monoxide (BrO). This compound will react with chlorine monoxide. The reaction forms molecular oxygen and frees the bromine and chlorine atoms, which then react with ozone again, repeating the process. Evidence for such a catalytic process comes from the 1987 observations by Willam H. Brune of Pennsylvania State University and Anderson of Harvard, who measured significant levels of bromine monoxide in the Antarctic ozone hole.

In addition to causing chemical reactions that convert the inert forms of chlorine to reactive forms (and the reactive forms of nitrogen into inert ones), PSCs also denitrify, or remove nitrogen from, the stratosphere. The slowly cooling water-ice clouds may be the primary agents that cause the denitrification. The cloud's water-ice particles not only form on nitric acid particles but also can absorb nitric acid in vapor form. The particles, which reach about 10 microns or larger, then fall from the Antarctic stratosphere as snow. This process both denitrifies and dehydrates the stratosphere. Nacreous clouds do not seem capable of removing nitrogen from the atmosphere. The rapid air currents through the cloud tend to evaporate the water-ice particles before any precipitation occurs.

Like the slowly cooling water-ice clouds, nitric acid trihydrate clouds seem able to denitrify the air as well. Their particles typically reach only about one micron in diameter—small enough to remain suspended. The small size results from the fact that the stratosphere contains very little nitric acid. Some clouds do form so slowly that their particles grow over one micron and thus may fall out of the stratosphere. Evidence for such a process comes from observations of the Arctic stratosphere, which has been denitrified but not dehydrated. In addition, lidar measurements by Browell show that some nitric acid trihydrate clouds contain particles larger than one micron.

The current PSC-chlorofluorocarbon theory for the formation of the ozone hole explains many observations. The release of CFCs from human activities, mainly in the Northern Hemisphere, is responsible for depleting ozone in the Southern Hemisphere because the long atmospheric lifetime of CFCs causes them to be uniformly distributed throughout the atmosphere. The ozone hole occurs near Antarctica during spring because the formation of the hole requires the presence of stratospheric clouds, which form only during the coldest times of the year. The first rays of spring sunlight initiate the chemical reactions that deplete ozone.

The ozone loss is more obvious over Antarctica than over the Arctic because the Antarctic stratosphere is colder, enabling more clouds to form, particularly below 20 kilometers. More clouds produce additional reactive chlorine atoms and remove nitrogen compounds, leading to greater ozone loss.

Perhaps a more important difference between the two poles concerns the longevity of the Antarctic vortex, a ring of rapidly circulating air that confines the ozone depletion. The vortex remains

intact throughout the polar winter, well into midspring. Ozone destruction begins in September with the return of sunlight. Loss of ozone reaches its peak in October. In the Arctic, where circulation patterns differ significantly from those in the Antarctic, the vortex has long since disintegrated by the time the polar spring (March and April) arrives.

A feedback mechanism may further extend the lifetime of the Antarctic vortex. Ozone absorbs sunlight, thus heating the atmosphere; depleted ozone levels cause the air to remain cold longer. Such cold air encourages the formation of PSCs and stabilizes the vortex. Measurements have revealed that over the past decade, the temperature in the vortex has declined, and the time that the vortex remains intact has increased. Lamont R. Poole of the NASA Langley Research Center and his colleagues have shown that PSCs form more frequently as the temperature drops, completing the positive reinforcement of the system.

The early breakup of the Arctic vortex makes it difficult to assess the magnitude of the ozone loss in the Northern Hemisphere. But measurements made in 1989 by Brune and Anderson show that nearly equal quantities of reactive chlorine occur in each polar vortex at altitudes of about 18 kilometers. Apparently, conditions exist in the Arctic that encourage the formation of an ozone hole.

Browell and Michael H. Proffitt of the NOAA Aeronomy Laboratory found in 1989 that large regions in the Arctic stratosphere above 18 kilometers suffered ozone depletion equivalent to 6 percent of the total amount of ozone over the Arctic. For comparison, total ozone loss over Antarctica averages 50 percent or higher. Greater ozone destruction did not occur, because regions at altitudes below 18 kilometers were too warm for PSCs to form. In future years, loss of ozone will increase as atmospheric levels of chlorine continue to rise.

Unless the Arctic temperature drops significantly, ozone depletion in the Arctic should never rival the loss in the Antarctic. A lower average winter temperature would enable PSCs to form over a region of greater altitude and to persist longer. An international scientific campaign is planned for this winter to help develop models of future Arctic ozone loss.

Because the ozone hole requires the presence of polar stratospheric clouds and a stable vortex for its creation, it is necessarily trapped near the poles, where few people live. That does not imply, however, the loss of ozone is restricted to the polar regions.

As the Antarctic vortex breaks up, pools of ozone-poor air spread over the Southern Hemisphere. In December of 1987 Rodger Atkinson of the Australian Bureau of Meteorology reported record-low ozone levels over Southern Australia and New Zealand after the breakup of the vortex. As those pools of low-ozone air spread out, they led to a small average ozone loss across the hemisphere. In addition, Adrian F. Tuck of the NOAA Aeronomy Laboratory has suggested that the Antarctic vortex may act as a chemical processor. Ozone-rich air may be entering the vortex, and ozone-poor air may be transported outward to the rest of the hemisphere. A high-altitude aircraft expedition being planned for 1993 will test this hypothesis and reveal more about the ozone depletion over Antarctica.

Not all ozone-destroying chemistry is confined to the polar regions. The sulfuric acid particles found throughout the stratosphere also trigger ozone reduction by freeing chlorine from its molecular reservoirs. Most significantly, those particles can also convert reactive nitrogen into inert forms, preventing the formation of chlorine reservoirs. The ozone loss caused by the sulfuric acid aerosols, however, is of a lesser magnitude than that caused by PSCs because of smaller mass and particle size.

Hofmann and Solomon have noted that the volcanic cloud created by the El Chichón eruption significantly reduced ozone levels. The amount of sulfate aerosols released rivaled the mass of nitric acid trihydrate PSCs. A large volcanic eruption early in the next century (when atmospheric chlorine levels will be higher) has the potential to induce significant global ozone loss.

Although many details remain unclear, investigators now understand quite well the fundamental link between stratospheric particulates and ozone destruction. Yet even the most thorough knowledge of stratospheric chemistry is unlikely to offer any quick technological means to stop the depletion. Replenishing the ozone lost annually—an amount comparable to the mass of the entire human population—would require at least 1 percent of the total U.S. energy output. This figure does not take into account the monumental task of transporting the ozone to the Antarctic vortex.

Instead the solution to ozone de-

pletion rests with international agreement among political leaders to restrict CFC production. In 1987 the U.S. and other industrial nations agreed to reduce the production of CFCs under the Montreal Protocol on Substances That Deplete the Ozone Layer. The protocol initially called for reducing emissions to 50 percent of 1986 levels before the year 2000. This reduction, however, would have allowed atmospheric levels of chlorine to increase to twice the current levels by the end of the next century.

As scientists became more certain of the chemistry of ozone depletion and the role that chlorine plays, the same nations recognized the need for swifter action. In the summer of 1990, they agreed to phase out chlorofluorocarbon production completely by the turn of the century.

Nevertheless, chlorine levels in the atmosphere will continue to increase over the coming decades. Large quantities of CFCs remain in refrigerators, air conditioners and foams, much of which will eventually be released into the atmosphere. Widespread use of safe replacements for CFCs seem to be at least a decade away.

Researchers predict that the amount of atmospheric chlorine will peak during the first decade of the next century. Because chlorofluorocarbons have such long lifetimes, chlorine may not return to levels that existed before the advent of the ozone hole until the middle of the next century, or even later.

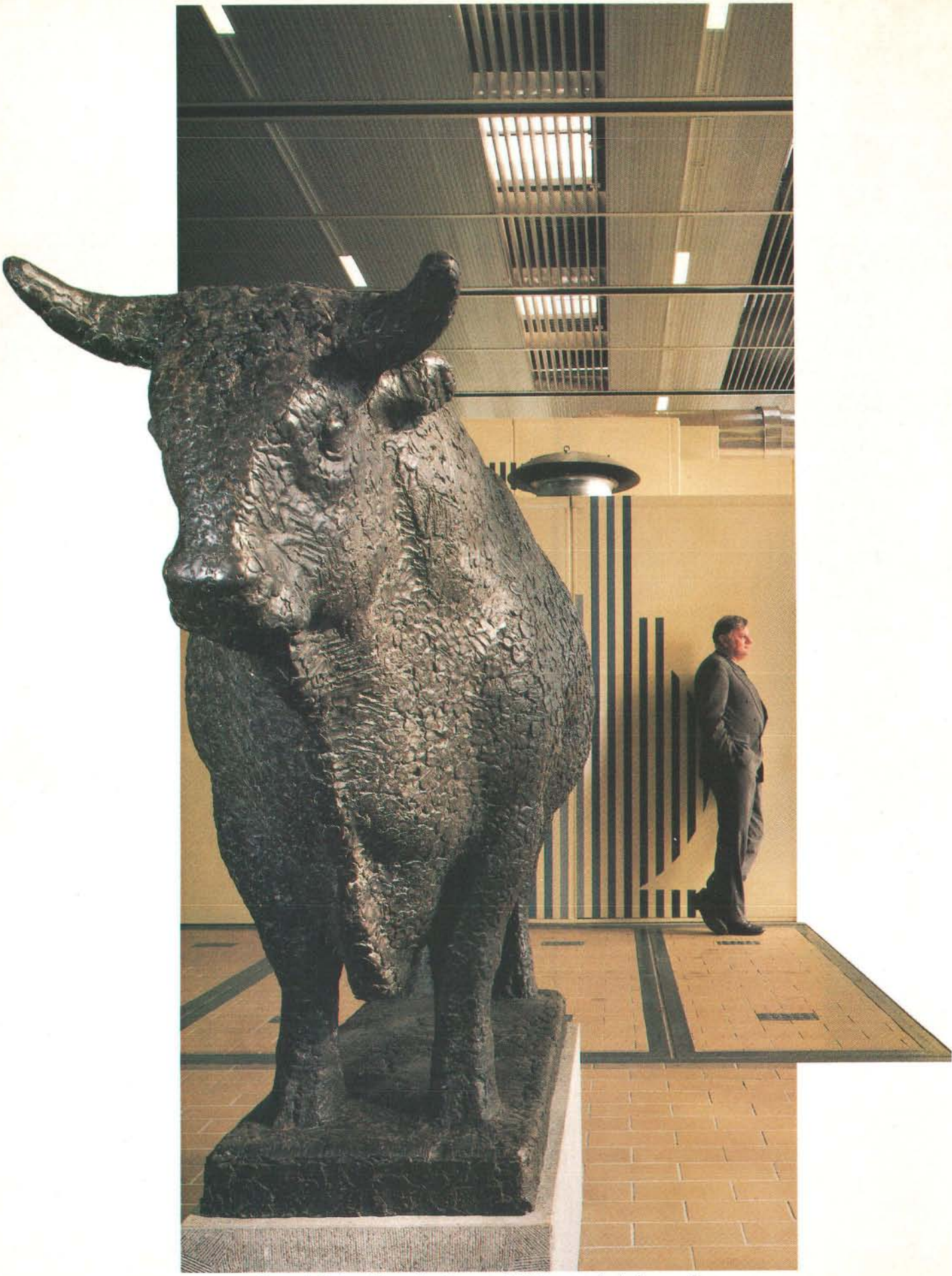
Consequently, the destruction of ozone will be more severe every year for the next few decades, leading perhaps to a doubling in area of the Antarctic ozone hole. In effect, society has wagered that somewhat greater ozone loss is less likely to disrupt ecosystems and human activities than are rapid control and disposal of CFCs. No one yet knows the odds of winning this bet.

FURTHER READING

PROGRESS TOWARDS A QUANTITATIVE UNDERSTANDING OF ANTARCTIC OZONE DEPLETION. Susan Solomon in *Nature*, Vol. 347, No. 6291, pages 347–354; September 27, 1990.

THE DYNAMICS OF THE STRATOSPHERIC POLAR VORTEX AND ITS RELATION TO SPRINGTIME OZONE DEPLETIONS. Mark R. Schoeberl and Dennis L. Hartmann in *Science*, Vol. 251, pages 46–52; January 4, 1991.

FREE RADICALS WITHIN THE ANTARCTIC VORTEX: THE ROLE OF CFCS IN ANTARCTIC OZONE LOSS. J. G. Anderson, D. W. Toohey and W. H. Brune in *Science*, Vol. 251, pages 39–46; January 4, 1991.



Frans de Ruiter, Managing Director of UNA, extends the beauty of art to an environment of high technology.

Frans de Ruiter's turbine is a work of art.

Frans de Ruiter is the Managing Director of UNA, the Dutch electrical power utility supplying the Noord Holland – Utrecht – Amsterdam area. He has a tough assignment. UNA serves one of the most densely populated regions of the most densely populated country in the world.

To help it do so, UNA has completed the installation of the two largest, most efficient gas turbines operational in the world today. Each generates 140 megawatts of electricity at efficiency levels well in excess of any comparable facility.

"We have made full use of ABB's most advanced technology to meet demanding targets," says Mr. de Ruiter. "At the same time, we have been able to satisfy Holland's strict environmental controls."

Not only are UNA's plants exceptionally "clean" – UNA has just won a prestigious international award for its environmental achievements – but the company's efforts to landscape the surroundings of its power plants have also won praise with local communities.

At the Utrecht power station, Mr. de Ruiter's environmental efforts have gone one step further. He has turned the interior of the plant into a giant gallery, and one of his new turbines itself into a work of art.

"Why not?" he asks. "Our employees deserve a stimulating work environment. And we are proud of having the world's most modern."

A hundred years of expertise in

- *Power Generation, Transmission and Distribution*
- *Industrial Automation* • *Transportation*
- *Environmental Systems*

makes ABB the world leader in electrical engineering.

ABB Asea Brown Boveri Ltd
Reader Services Center
P.O. Box 822
CH-8021 Zurich, Switzerland

ABB
ASEA BROWN BOVERI

Early Bow Design and Construction

The most effective tool of hunter and warrior for millennia, early bows display a variety of modifications that reflect the functional requirements of their users

by Edward McEwen, Robert L. Miller and Christopher A. Bergman

Few would dispute that the invention of the wheel and the mastering of fire rank among the most important developments in history. Sometimes overlooked, however, is the creation of the bow. From Paleolithic times until the advent of firearms in the 16th century, the bow was not only a major hunting tool but also the primary weapon in combat. The bow proved vital to the many central Asian nomads who conquered lands and founded dynasties in China, as well as to the attackers laying siege to the medieval castle in Europe. Found in virtually all cultures, bows display important and practical variations in construction, ranging from designs barely more than branches with string attached to what can only be described as sophisticated mechanical devices.

Fundamentally, a bow is a two-armed spring spanned and held under tension by a string. Drawing the bow places the back, or outside curve, under tensile stress and the belly, or inside curve, under compressive forces. Any bow must adapt to these forces to avoid breaking and to propel the arrow successfully. When fully drawn, the bow stores potential energy in its limbs. Releasing the bowstring transfers this energy to the arrow, throwing it into flight.

Immortalized in legend and in history, the English longbow is perhaps the most familiar bow. Yet for all its fatal power, the longbow would not be very practical to shoot while, for instance, riding a horse. From archaeological excavations and our own experiments with replicas, we know that early bowyers adapted bows to a particular need and so produced myriad subtle variations in design. Some peoples, such as the Teton Lakota (Sioux), shortened their bows to ease handling on horseback. Others, such as the Huns, combined different materials to make small but exceptionally powerful bows that could send an arrow through the metal body armor of an enemy.

From the Paleolithic era onward, bow design generally followed separate trends in Europe and Asia. Neither path can be considered intrinsically better than the other. Rather, every bow design represents one possible solution to the problem of hurling a small, lightweight projectile with accuracy and penetrating power.

The various kinds of bows did not appear suddenly. Bow design seems to be part of a gradual process of modification, spanning many millennia and prehistoric cultures. For instance, some scholars believed that the medieval Anglo-Saxons, the Normans or the Welsh invented the English longbow. But in fact, investigators have found antecedents that date back at least 8,000 years. Some evidence hints that the first archery tackle appeared during the earlier Upper Paleolithic (circa 35,000 to 8000 B.C.). In this article, we trace the evolution of the bow from its prehistoric beginnings to the modifications introduced, primarily in Europe and Asia, as recently as 400 years ago.

The earliest evidence for the origin of the bow may be projectile points recovered from Old World Paleolithic societies, such as the Perigordian and Solutrean cultures in what is now France. The thin, narrow basal parts of these points could easily fit into a narrow slot at the end of an arrow shaft. It is equally likely, however, that the points, which date to between 28,000 and 17,000 B.C., were used instead to tip a thrown dart.

Archaeologists have gathered more positive evidence at Stellmoor, near Hamburg, Germany, recovering a number of wooden arrow shafts and foreshafts from a late glacial culture that existed in the early ninth millennium B.C. There can be no doubt that the broken shafts were indeed for use with a bow. Unlike thrown darts, which have a narrow depression, or cup, that fits into the hook of a device called a spear thrower, these shafts have shallow, rectangular slots, or nocks, that could fit only a bowstring.

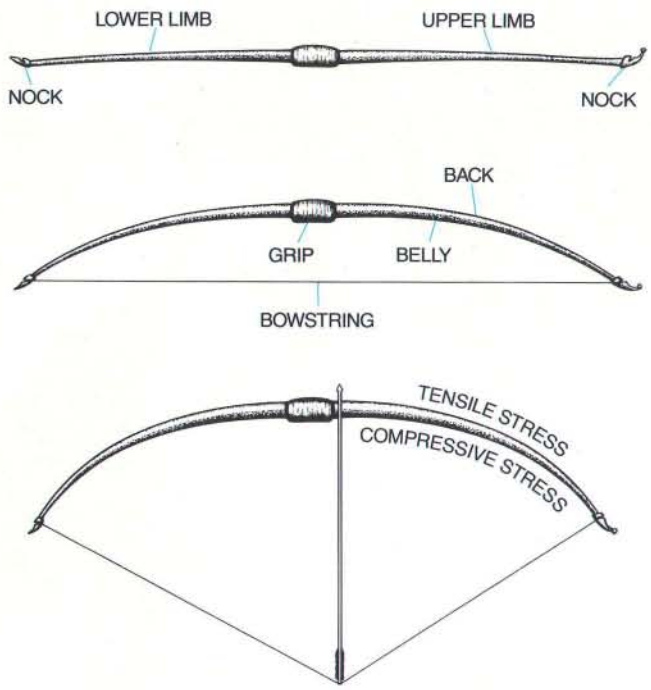
The earliest complete bows investigators have recovered date to around 6000 B.C. Preserved in waterlogged regions of Scandinavia, the bows are simple, made of one complete piece of wood, primarily yew or elm. Because they consist of a single raw material,

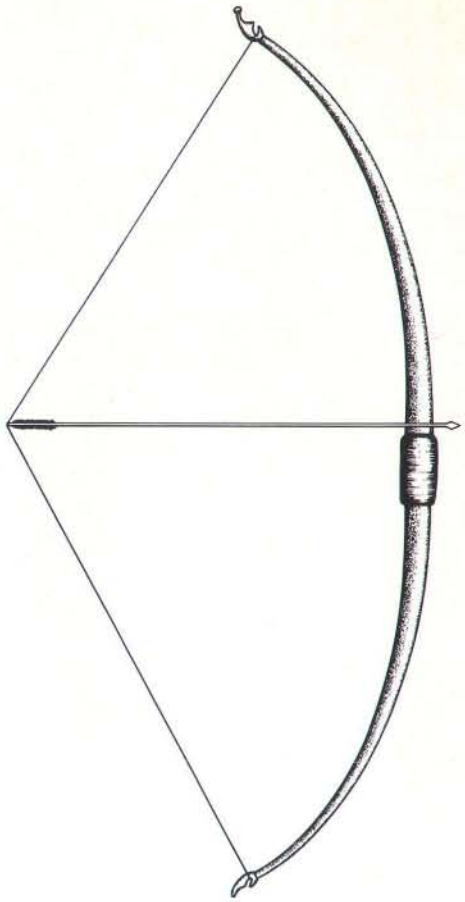
MANCHU ARCHERS, armed with composite bows, fight Tatar warriors in this copperplate engraving from 1765. Commissioned by Emperor Ch'ien-lung, the scene celebrates the conquest of the Tatars in the mid-18th century.

EDWARD McEWEN, ROBERT L. MILLER and CHRISTOPHER A. BERGMAN have studied and experimented with the bows of cultures from around the world. McEwen is currently editor of the *Journal of the Society of Archer-Antiquaries*. He has studied and translated medieval Persian archery texts at the School of Oriental and African Studies in London. Miller is an archaeologist with Clover Archaeological Services in Northport, N.Y. His research applies scientific techniques to the study of early technology, human ecology and paleoepidemiology. Bergman is principal archaeologist with 3D/Environmental Services, Inc., in Cincinnati and has been studying and replicating Native American archery tackle for the past 10 years. Miller and Bergman received their doctorates from the Institute of Archaeology in London. The authors wish to thank Charles E. Grayson of Clatskanie, Ore., and Frank J. McAvinchey of 3D/Environmental Services for their help with this article.

The Self-Bow

Lengthening of the limbs, a design technique most familiar in the English longbow, improves performance by increasing draw length.





they are termed self-bows, along with other members of this class.

By the Mesolithic period (circa 8000 to 3200 B.C.), a sophisticated bow design had emerged in northern Europe. For example, the bows recovered from the Holmgaard bog on the island of Zealand, Denmark, are made from single staves of elm. A rigid grip area constricts wide, flattened limbs that gradually become narrower toward the tips. The early bow makers must have carefully scraped and thinned the belly (a process called tillering), for the bows display an even curve when strung. All these characteristics contribute to a uniform distribution of stress along the length of the bow, thereby reducing the chance of breakage and improving performance.

The designers of the Holmgaard bows also made them long, about 150 to 180 centimeters—roughly the size of the medieval longbow. A longer bow allows for an increased draw length, which contributes significantly to the speed of the shot and to the cast, or the distance the bow will shoot an arrow. Short self-bows, such as those used by mounted Lakota and Comanche on the plains of North America, have shorter draw lengths, often only 55 to 60 centimeters.

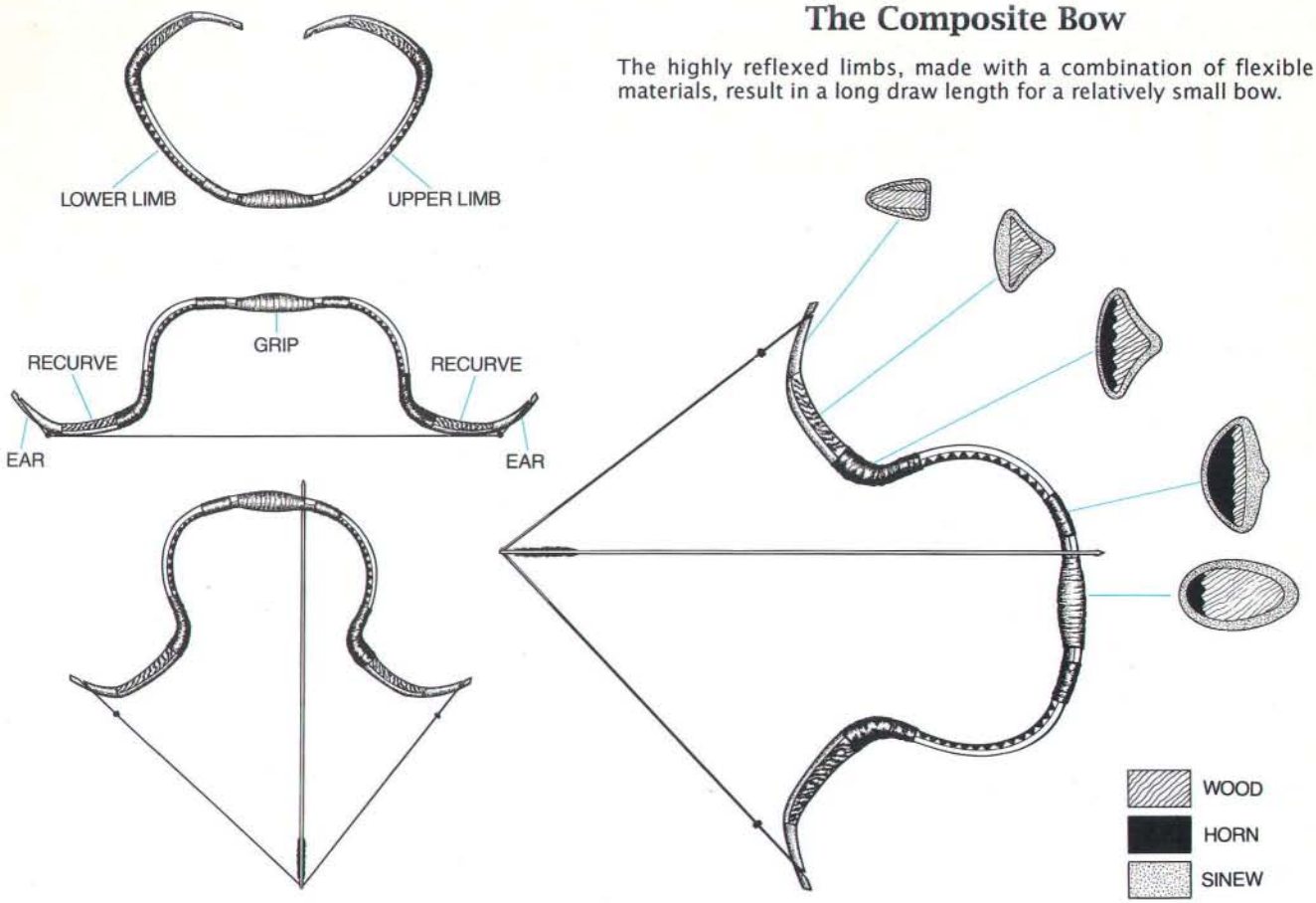
The stages leading to the development of the Holmgaard bows must have depended on simple experiment and on understanding the limitations imposed by the raw materials and the tools used in manufacture. Neolithic self-bows perhaps best illustrate the role that available tools and materials play in bow design. Using the same repertoire of stone implements available to Neolithic bowyers, we found the tools quite capable of constructing self-bows. Of course, such implements limit the sophistication and degree of the woodworking. For example, the Neolithic Meare Heath yew longbow from Somerset, England, radiocarbon dated to within 120 years of 2690 B.C., differs substantially from the familiar medieval longbow (made with metal tools) in the shape of its back and belly. Although both are similar in size (about 200 centimeters), the Meare Heath longbow has a more rounded, convex back and a flatter belly than its more modern counterpart.

The design indicates that the makers of the Meare Heath bow exploited the shape of the original tree branch or trunk to reduce manufacturing time. We believe the prehistoric bowyer first carefully selected and then split a suitable branch or mature sapling of the desired width and length. The actual woodworking must have consisted primarily of tapering the stave in width and thickness. The modest amount of woodworking is especially evident on the back, which consists largely of the outer, natural curve of the original stave. A rounded back disturbs the layers of wood beneath the bark as little as possible, making it less likely for the bowyer to weaken the structure inadvertently by, for example, cutting across the grain. Most bows break because tensile stress causes failure of the back. Frequently, the weak point is where the grain has not been carefully followed; the wood fibers separate, cracking the bow.

Some confusion exists among scholars concerning the material composition of some of the early yew bows. Yew is composed of two visibly distinct layers of wood: the white-colored sapwood, which is the physiologically active, outer layer of wood, and the orange-red heartwood, which is the nonliving, central part. Sapwood is elastic

The Composite Bow

The highly reflexed limbs, made with a combination of flexible materials, result in a long draw length for a relatively small bow.



and strong under tension; heartwood is better suited to handle compressive stress. Gad Rausing of Lund University noted the apparent absence of sapwood on the backs of Neolithic yew bows from lakeside sites in Switzerland, dated between the fourth and third millennia B.C. Researchers have also reported the absence of sapwood in the Meare Heath yew bow.

Our experience in making yew bows indicates that it is highly unlikely that these Neolithic weapons were made of heartwood alone. Yew heartwood is simply too brittle to resist the high tensile stress associated with stringing and drawing a bow. A weapon made of heartwood alone would be quite unreliable and prone to break at any time. It may be that the Neolithic examples were originally made from unseasoned wood, which may have been more elastic, but the performance of such a bow would have been poor at best.

The rise of metal tools after 2000 B.C. enabled bowyers to modify the longbow differently. With many well-preserved examples, the medieval longbow perhaps best illustrates the kinds of modifications produced by metal tools. The English bowyers made their longbows from staves split from trees that were larger and more mature than those used by their Neolithic counterparts. Because the curvature on the outside of a large tree was wider than that of a branch, the later bowyers could produce a bow with a flatter back. For example, the bows recovered from Henry VIII's warship *Mary Rose,* which sank on July 19, 1545, were rounded in cross section, with a slightly flattened, sapwood back.

Modifications during Victorian times tended to focus on the thickness of the stave rather than its width. This emphasis creates the highly "stacked" section so common in these bows. A highly stacked bow produces a faster and longer cast for a lighter draw weight. Yet the uneven distribution of forces along the narrow center line of the thick, rounded belly disposes such a bow to failure.

These variations do not necessarily reflect progressive evolution. Indeed, "ideal" bow designs were mathematically developed and scientifically tested in the 1930s and 1940s by Clarence Hickman of Bell Laboratories, Forrest Nagler of the American Society of Mechanical Engineers and Paul Klopsteg of the Ordnance Department of the U.S. Army. These designs correspond more closely to the wide-limbed flat bows used in Mesolithic and Neolithic Europe rather than to later bows developed in England. No one knows why the English chose narrow limbs. It may have been done to maximize the amount of raw material at hand.

Outside Europe, bow design embarked on a different route. Although the simple self-bow undoubtedly arose independently a number of times in many cultures, the most complex modifications occurred in Asia. Unlike the Europeans, the Asians seemed to concentrate not on the architecture of the limbs but on the materials. In particular, Asian bowyers used adhesives derived from hide and fish swim bladder to glue animal sinew to the backs of their bows.

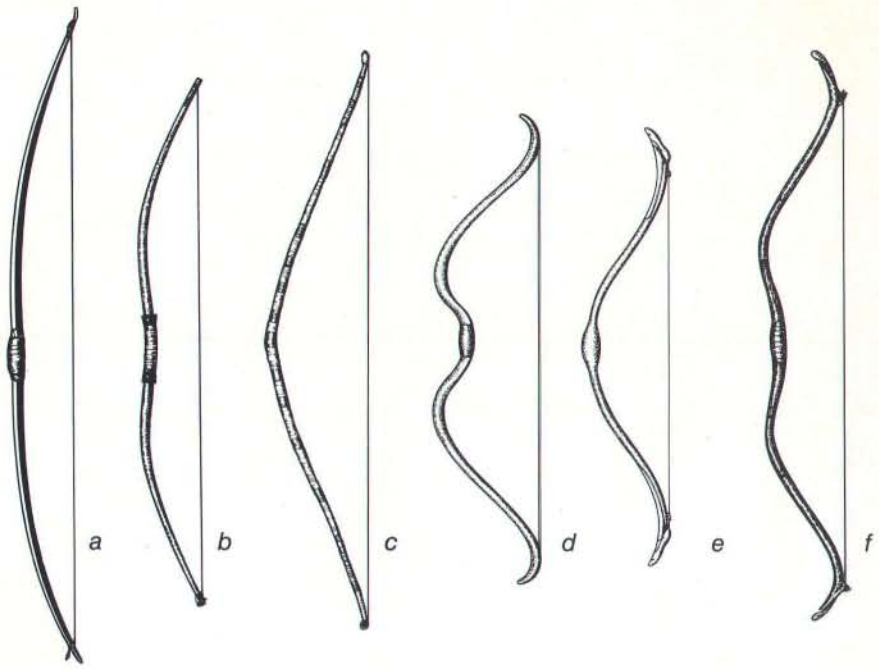
Sinew has high tensile strength, estimated at about 20 kilograms per square millimeter, roughly four times that of bow woods. Such strength enables a bow to be shortened significantly without loss of draw length or increased risk of breakage. Easy to handle on horseback, these short-limbed, sinew-rein-

forced bows were used in northern Asia and the Far East. Some Indian tribes on the plains of western North America also developed and used such bows. (The horse is not necessarily a prerequisite to the development of sinew-reinforced bows; Native Americans in California hunted on foot with them in forests.)

A chief advantage of sinew-reinforced bows is that they are invariably reflexed: the limbs of the unstrung bow reverse themselves. Reflex places the limbs under greater tension when the bow is strung, storing more power than self-bows can. Shorter limbs also mean a more efficient transfer of energy. The long, heavy limbs of high-power self-bows use a great deal of energy as they move forward with the release of the string, resulting in a diminished and inefficient transfer of energy.

Early bowyers in eastern and western Asia did not stop at simply adding sinew to the bow. Some must have recognized that other materials exist in nature that are stronger than wood. They created the most sophisticated bow, which required a significant level of skill to produce. Called the composite bow, it is a mechanical tour de force. As the name suggests, this type of bow combines several different materials. In its classic form, it consists of a thin wooden core with sinew glued to the back, and horn, usually from the water buffalo, glued to the belly. Modern archers have given this type of bow other names, including laminated, backed or reinforced, and compound. We use the term "composite" here to refer to the fully developed bow made of horn, wood and sinew.

The composite bow exploits the materials used in its construction. The sinew on the back handles tensile stress. The horn, with a maximum strength of roughly 13 kilograms per square millimeter (about twice that of hardwoods), bears compressive loads. Horn also has a high coefficient of restitution, or the ability to return to its original shape after being distorted. The flexibility of these materials gives the bow short, lightweight, reflexed limbs capable of storing a large amount of energy under tension. In addition, the flexible limbs enable the composite bow to be drawn much farther relative to the overall length of the weapon. The combination of extended draw length and short limbs enables the composite bow to shoot an arrow faster and farther than can a wooden self-bow of equal draw weight. We conducted tests to show that a replica composite bow with a draw weight of 27 kilograms will shoot the same arrow as fast as a replica medieval yew longbow with a draw weight of 36 kilograms (about 50 meters per second).



Representative Bows

The fundamental types of bows are illustrated by the medieval yew longbow (*a*), Teton Lakota sinew-reinforced bow (*b*) and four kinds of composite bows: western Asian angular bow (*c*), Scythian bow (*d*), 17th-century Turkish bow (*e*) and 17th-century Crimean Tatar bow (*f*).

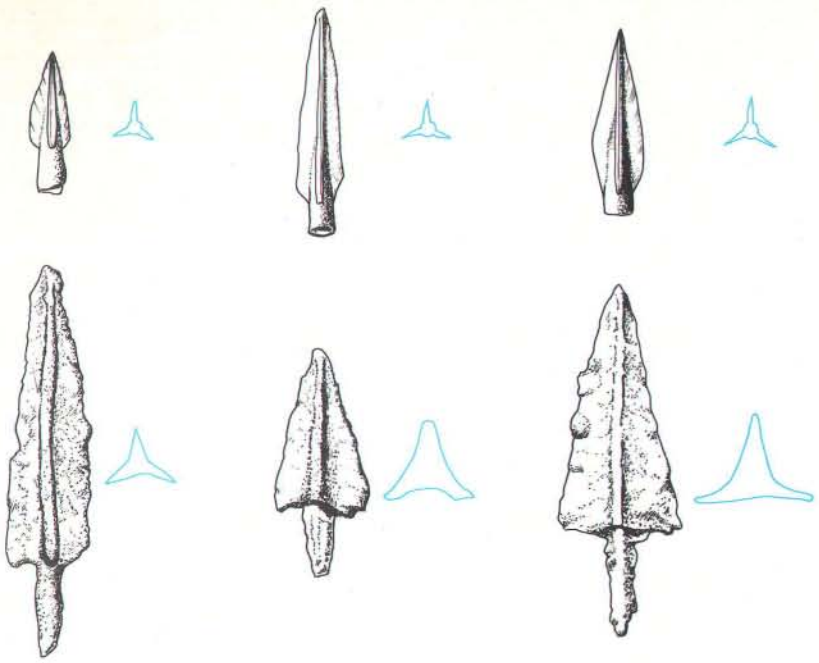
Only the crossbow, invented around 500 B.C., can propel a projectile faster and farther [see "The Crossbow," by Vernard Foley, George Palmer and Werner Soedel; SCIENTIFIC AMERICAN, January 1985]. The crossbow, however, is inefficient; to achieve enormous draw weights, which reach up to one ton, the crossbow requires mechanical parts. Thus, it is not a fair comparison for an ordinary hand-shot bow.

Another advantage of the composite bow is that it can be kept strung for prolonged periods without adverse effects. Simple wooden self-bows and sinew-reinforced bows are usually kept unstrung between uses to avoid "string follow" and a loss of power.

We do not know precisely where or what people first invented the composite bow. The archaeological and historical record suggest that various cultures independently developed the composite bow during the third millennium B.C. Specifically, current evidence indicates that the composite bow developed separately but simultaneously in Mesopotamia and Anatolia and in the steppes of northern Asia.

The English general and archaeologist Augustus Henry Pitt-Rivers, who first introduced the term "composite bow" in the late 19th century, considered these bows to be the product of peoples living in areas where no suitable wood existed for bow manufacture. At first glance, Pitt-Rivers's premise seems logical. But in fact, the composite bow first appears in the archaeological record in areas with ample supplies of good bow wood. The ancient Egyptians, for example, made and used the composite bow, but they also produced self-bows from hardwoods such as acacia and carob.

If the lack of suitable wood is not the reason for the development of the composite bow, then one must assume that it originated from the desire to produce a mechanically superior weapon. The process leading to its invention may be related to an increase in horse transport during the third millennium B.C. in Asia, when the use of chariots and mounted cavalry in warfare became widespread. The Asian horsemen most likely preferred a shorter bow



Early Arrowheads

Archery tackle helps archaeologists deduce the requirements of their users. For instance, Scythians made bronze arrowheads, shown here from the third century B.C., about 25 to 50 millimeters long (*upper row*). The development of body armor, however, necessitated heavier and larger arrowheads made from iron, which could penetrate armor. The examples are those used by the Huns (*lower row*). The outline to the right of each arrowhead shows its shape when viewed head-on.

and increased its power and reliability by strengthening it with other materials. The evolution of the composite bow in early Asia probably mirrors bow design in North America after the 16th century. The mounted Plains Indians experimented with bow design by gluing sinew to the backs of their weapons. Many tribes later removed the wood entirely, substituting elk antler or mountain sheep horn for the wood belly—a development only one stage away from a true composite bow.

One of the earliest surviving examples of the composite bow is the western Asian angular bow, which appeared during the third millennium B.C. This bow forms a shallow triangle when strung and a semicircle at full draw. Illustrations of these bows appear in Mesopotamian seals, on Egyptian tomb paintings and on Assyrian monumental reliefs, indicating that this design was used for nearly 2,000 years, from 2400 B.C. to about 600 B.C.

Aside from artistic representations, archaeologists have found numerous examples of angular composite bows in funerary chambers in Egypt. In 1922 Howard Carter recovered the most well-known material from the tomb of Tutankhamen. The tomb contained 32 angular composite bows, 14 wooden self-bows and 430 arrows, as well as quivers and bow cases.

Initial theories regarding the function of the angular composite bow could not account for the apparent ability of the bow to bend through the handle. This action was entirely in contrast to the design of the traditional longbow, which would "kick," or jar, in the hand if not for the stiff grip section.

Our replication of the angular composite bow has demonstrated that its center section only gives the appearance of bending. The angle at the center of the bow is inflexible, and the bow actually bends throughout the length of the limbs. Releasing the bowstring produces no kick, which results in a smooth, accurate shot. The extremely long draw length, reaching 101 centimeters with the limbs under maximum tension, would have provided a greatly enhanced cast compared with that of the second-millennium B.C. self-bows.

The angular bow held sway in western Asia until the late seventh century B.C., when the Scythians joined the Medes and Babylonians to bring the Assyrian Empire to an end. The Scythians, noted horsemen and archers, apparently originated in the steppes of eastern Ukraine in modern Russia. A nomadic people, they covered vast areas of Asia and left examples of their distinctive small, bronze trilobate (three-lobed) arrowheads, averaging between 25 and 50 millimeters, from China to Greece.

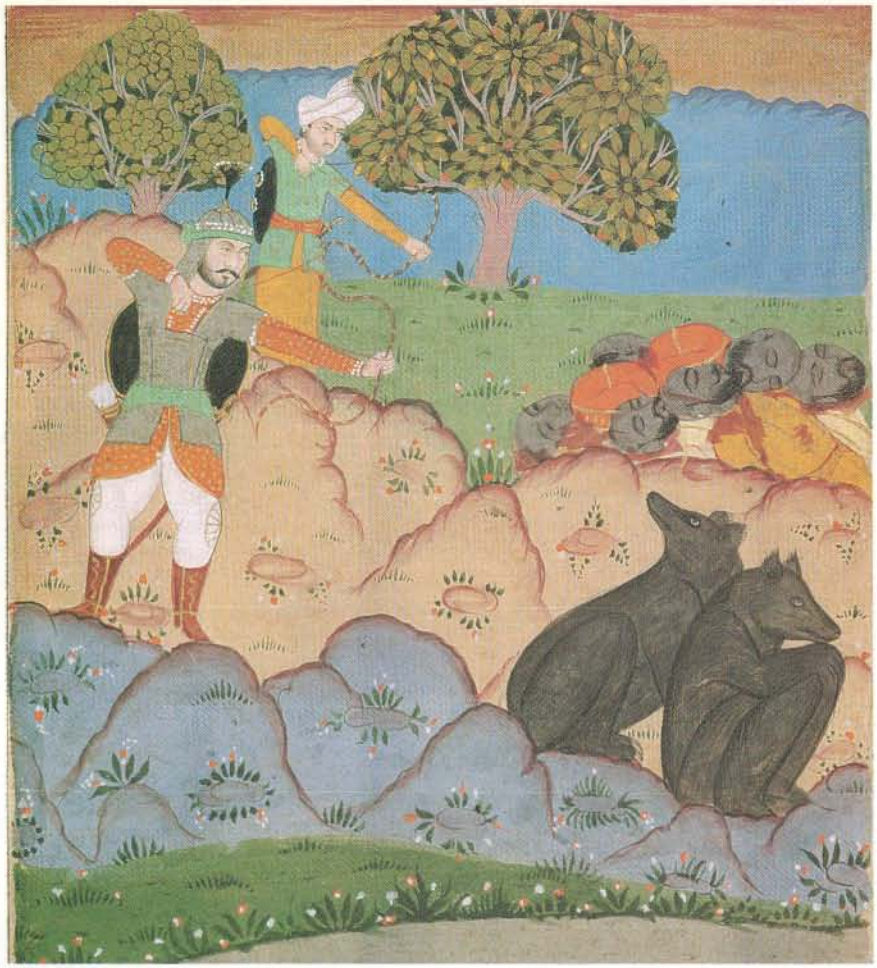
Most data collected on Scythian archery tackle derive from artistic representations. In addition, a number of bow cases and quivers, as well as arrow shafts, have been recovered from burials at Pazyryk in the eastern Altai region of Soviet central Asia. Although we must always treat artistic representations with caution, the consistent uniformity in the depiction of the Scythian bow enables us to draw some conclusions.

The design type commonly referred to as the Scythian bow, or *scythicus arcus* of the Romans, was used by many different peoples over a long period. The bow was fully developed as early as the ninth century B.C. by the Cimmerians, from north of the Caucasus. Later, the Scythians introduced it to the ancient Greeks. The design eventually reached as far as northern France.

Contemporary illustrations and measurements of bow cases recovered at the Pazyryk site indicate that the Scythian bow was about 127 centimeters long. Shaped like Cupid's bow, it had a set-back handle and reflexed limbs terminating in recurved ends. Such a length and design, incorporating a heavy reflex in the handle section and flexible limbs, provide a draw length of about 76 centimeters.

This figure agrees with the length of the arrows recovered from the Pazyryk burials. Some scholars estimate the draw length at only 45 centimeters, based on representations from fifth-century B.C. Attic vases and on the small diameter of the socket at the base of the trilobate arrowheads. The small size of the Scythian arrowhead and its socket no doubt suggested that the arrow shaft was correspondingly reduced in length and diameter.

But this deduction fails to take account of the chief advantage of the composite bow: its high ratio of draw length to bow length. Thus, a relatively short, 127-centimeter bow could be drawn much farther than its actual length suggests. The bowyers probably made the base of the trilobate arrowhead narrow to fit into a barreled arrow shaft. This design, used into medieval times, was tapered, with the thickest point in the middle of the shaft. Like the angular



PERSIANS WITH COMPOSITE BOWS hunt down what are apparently werewolves. The mythical scene is one of several in an 18th-century book detailing the life of a Persian prince. The names of the prince, artist and author remain unknown.

composite bow, the Scythian bow appears to have been totally flexible. Its limbs had none of the positive stiffening later composite bows attained because bone or antler plates were added to the grip and the nock ends, or ears.

A weapon often evolves simultaneously with attempts to produce more reliable protection against it. In the third century B.C. the eastern neighbors of the Scythians developed new techniques of warfare. Called the Sarmatians, they covered their cavalrymen and horses with armor and trained them to fight in close formation. The tough body armor necessitated the development of a bow that could propel an arrow with a heavy iron tip faster and with greater impact.

Central Asian nomadic peoples, such as the Huns and Avars, provided the means to penetrate the armor. They stiffened the ears of the bows and set them at a sharply recurved angle. The changes formed a compound lever at the end of each limb. Such levers enable the archer to bend a heavier bow limb with less effort; the angle at which the ear curves toward the back of the bow makes it act as if a large-diameter wheel were attached at the end of each limb.

As the archer draws the bow, the "wheel" unrolls, in effect lengthening the bowstring. Releasing the bowstring moves the ears forward and effectively shortens the string, providing an increased acceleration to the arrow. This development precedes by many centuries the modern compound bow, which uses a system of pulleys to achieve a similar but more marked effect.

By the 17th century, peoples such as the Ottoman Turks and the Turkish tribes of Iran had modified the basic composite bow design of other Asian nomads. Turkish bowyers began to experiment with bows shortened to around 111 to 116 centimeters. They dispensed with the set-back handle and bone or antler plates used for the ears of earlier weapons. The result was a bow that gracefully curved on either side of a rigid handle toward slightly recurved ends.

These short bows had an extended draw length and were tremendously powerful. Their draw weights ranged from 36 kilograms to more than 45 kilograms. Such draw weights are equivalent to those of the English longbow, which is almost twice as long. Armed with the Turkish bow, the mounted Ottoman cavalry proved formidable, conquering eastern Europe during the Middle Ages. But with the invention of gunpowder and muskets, the Turkish bow gradually declined in use as a military weapon; it became favored instead for sport—especially flight shooting. Perhaps the finest shot ever recorded was taken in 1798 by the Ottoman Emperor Sultan Selim III, who shot two flight arrows 889 meters, a feat never equaled with archery tackle made along traditional designs.

Archery tackle, like any other utilitarian equipment, reflects the nature of the material available as well as the functional requirements of its users. For example, the large, heavy composite military bows built by the Crimean Tatars of the 17th century make the tiny bows of modern Kalahari Bushmen look like toys. Bushmen bows, however, function extremely well within their proper context—that is, they can catapult a small, unfeathered arrow that penetrates, leaving poison beneath the skin of an animal.

In short, the effectiveness of a bow design can be gauged only by its ability to function successfully within the context for which it was created. For many thousands of years, the bow remained the most effective missile-firing agent available to hunters and warriors. Its supremacy began waning only in the 16th century with the invention and widespread distribution of a more powerful weapon yet: the firearm.

FURTHER READING

NEOLITHIC BOWS FROM SOMERSET, ENGLAND, AND THE PREHISTORY OF ARCHERY IN NORTH-WEST EUROPE. J.G.D. Clark in *Proceedings of the Prehistoric Society,* Vol. 29, pages 50–98; December 1963.

THE BOW: SOME NOTES ON ITS ORIGIN AND DEVELOPMENT. Gad Rausing. Bonn, Rudolf Habelt, and Lund, CWK Gleerups, 1967.

TURKISH ARCHERY AND THE COMPOSITE BOW. Third Edition. Paul E. Klopsteg. Manchester, Simon Archery Foundation, 1987.

EXPERIMENTAL ARCHERY: PROJECTILE VELOCITIES AND COMPARISON OF BOW PERFORMANCES. C. A. Bergman, E. McEwen and R. Miller in *Antiquity,* Vol. 62, No. 237, pages 658–670; December 1988.

SCIENTIFIC AMERICAN

PRESENTS A SPECIAL ISSUE

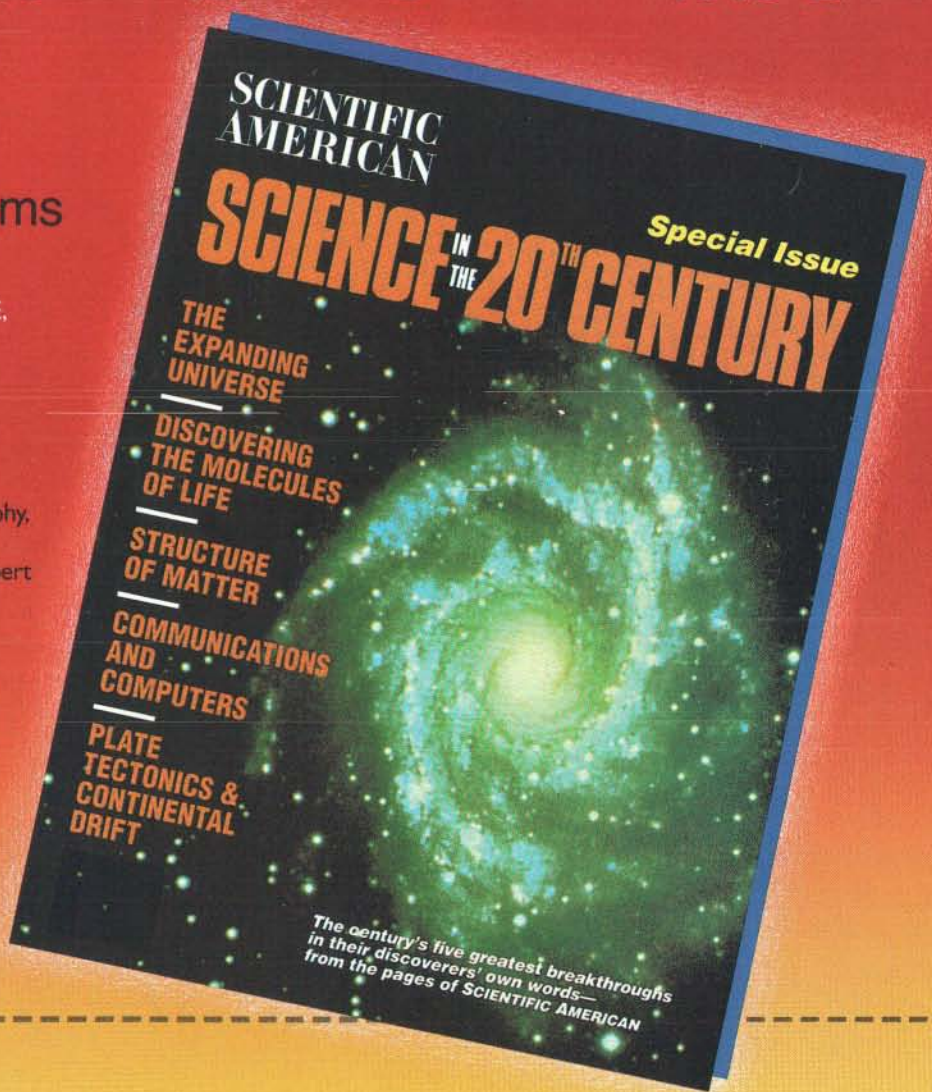
Sponsored by

AT&T Network Systems

THE FIVE MAJOR ADVANCES OF SCIENCE, AS REPORTED IN THE PAGES OF *SCIENTIFIC AMERICAN*, IN THE ACTUAL WORDS OF THE SCIENTISTS WHO MADE THEM HAPPEN.

- Introduction by Jonathan Piel, Editor of SCIENTIFIC AMERICAN
- Extensively illustrated: elegant color photography, art, and graphics
- Articles by Albert Einstein, J. Tuzo Wilson, Robert Weinberg, Steven Weinberg, F.H.C. Crick, Alan H. Guth, Paul J. Steinhardt and others
- Excellent classroom teaching aid
- Over 175 pages, never before collected in one publication

The discoveries reported by these authors are both pivotal and fascinating. The inflationary universe, the structure of genes, the essence of matter, transistors, lasers, the earth's hot spots, and more—all major scientific breakthroughs in the 20th Century and each one revolutionizing society, the economy, the way we know ourselves. Authoritatively written by the scientists themselves.



☑ YES, I would like to receive the special issue SCIENTIFIC AMERICAN *Science in the 20th Century* as soon as it is available.

Please send: _____ copies x $3.95 $_____
$4.95 outside the U.S.
Add $1.00 per copy postage and handling. $_____
TOTAL ENCLOSED $_____

Postage & handling paid by SCIENTIFIC AMERICAN on orders of ten or more copies.

Name _____

Company _____

Address _____ Apt. _____

City _____ State _____ Zip _____

STC 6

ON SALE NOW AT YOUR NEWSSTAND

Please make check or money order payable to: SCIENTIFIC AMERICAN.

Send order to:
SCIENTIFIC AMERICAN
Dept. 20C
415 Madison Avenue
New York, NY 10017
U.S.A.

Laser Surgery

Precise, powerful and at times subtle in their effects, lasers are increasingly important medical tools. These knives of light can be used to treat individual cells as well as whole organs

by Michael W. Berns

Lasers make good scalpels. Rather than having to slice through everything they encounter, these instruments can be highly selective. That specificity allows lasers to penetrate to the interior of a cell, or an organ, while leaving the exterior intact—something no surgeon's knife can do.

The refinement of this precision over the past three decades has enabled the medical uses for lasers to proliferate. Initially, the heat generated by the laser beam was harnessed to destroy tissue. Now, although the thermal effects are still most commonly used in medical procedures, other, nonthermal effects are also proving valuable for both treatment and diagnosis. In addition to heating tissue, photons from laser beams can drive chemical reactions, break the atomic bonds that hold molecules together or create shock waves.

Indeed, biomedical applications for lasers include such diverse tasks as unclogging obstructed arteries, breaking up kidney stones, clearing cataracts and even altering genetic material. Lasers can also provide information about the inner workings of cells. The biological understandings gained from such studies may have medical implications.

Perhaps because of its promise, laser technology has at times been clouded by hyperbole. The realistic future of laser surgery rests on attaining a fuller appreciation of the basic physical and chemical mechanisms by which light interacts with organs and organelles. These insights will establish when it is right to use the laser and when it is not. Recognizing that a scalpel or a $500 electrocautery knife can be more appropriate and less risky than a $100,000 laser device is pivotal to the continued success of this technology in medicine.

MICHAEL W. BERNS is the Arnold and Mabel Beckman Professor at the University of California, Irvine. He received his undergraduate and graduate degrees from Cornell University and joined the faculty at Irvine in 1972. Ten years later Berns and industrialist Arnold O. Beckman founded the Beckman Laser Institute, an academic center where clinicians, biologists, engineers and physicists collaborate on laser research. (Their subjects range from chromosomes and cells to tigers and humans.) Berns dedicates this article to Arnold O. Beckman in celebration of his 91st birthday.

The idea of using light for surgery antedates the laser. In 1946 a German physician, Gerd Meyer-Schwickerath, used the sun to treat detached retinas and to destroy tumors in some of his patients' eyes. In 1961, only a year after Theodore H. Maiman built the first laser at Hughes Aircraft Research Laboratories, Milton Zaret of the New York University School of Medicine used a laser to produce ocular lesions in animals. Human trials followed two years later, when Chris Zweng of the Palo Alto Medical Research Foundation in California treated retinal disease in his patients. The laser was quickly accepted as a standard ophthalmic surgical tool.

The pioneers of laser surgery used the beam of light because of the intense heat it generated. Today most laser surgery makes use of this heat, primarily because its destructive effects can be extremely selective and precisely controlled. If the wavelength of light from the laser is matched very closely with the absorption band of the target structure, the laser light will be absorbed by, and therefore damage, that structure.

The dark brown melanin pigment of the retina, for instance, absorbs the green beam of the argon laser. Consequently, the argon laser can destroy specific regions of the retina without harming other areas of the eye, which absorb different wavelengths of light. The procedure can effectively treat diabetic retinopathy, a degenerative disease that used to account for a large proportion of acquired blindness in the U.S.

Red birthmarks, called port-wine stains, also absorb the argon laser's beam, which can be blue or green, depending on its wavelength. The light destroys the hundreds of extra blood vessels that lie just beneath the skin's outer layer and discolor it. Although in this case laser surgery is preferable to skin grafting and incision, it has its own disadvantages. The heat generated by the beam can sometimes spread to parts of the skin other than the abnormal blood vessels and cause scarring, or loss of pigment.

Avoiding such extensive damage led to an important advance in laser surgery. In 1983 R. Rox Anderson and John A. Parrish of Harvard University suggested that short exposure—less than one one-thousandth of a second—to intense light would destroy the absorption site but produce little or no damage to the adjacent tissue. They reasoned that it should take less time for the energy to be absorbed and the subsequent heat to be dissipated than it does for the heat to be transferred to surrounding areas. Therefore, the selective destruction of pigmented targets would require two conditions: preferential light absorption and sufficiently short light pulsation.

That theory proved to be true. Selective photothermolysis, as the technique is called, has indeed improved the treatment of port-wine stains. It has also proved useful for removing tattoos. Scarring can be prevented when the laser beam is delivered in short pulses rather than continuously or in long pulses, which last roughly a quarter of a second. (Many lasers could ideally be delivered either continuously for thermal effects extending beyond the absorption site or in short pulses for destruction confined to the target.)

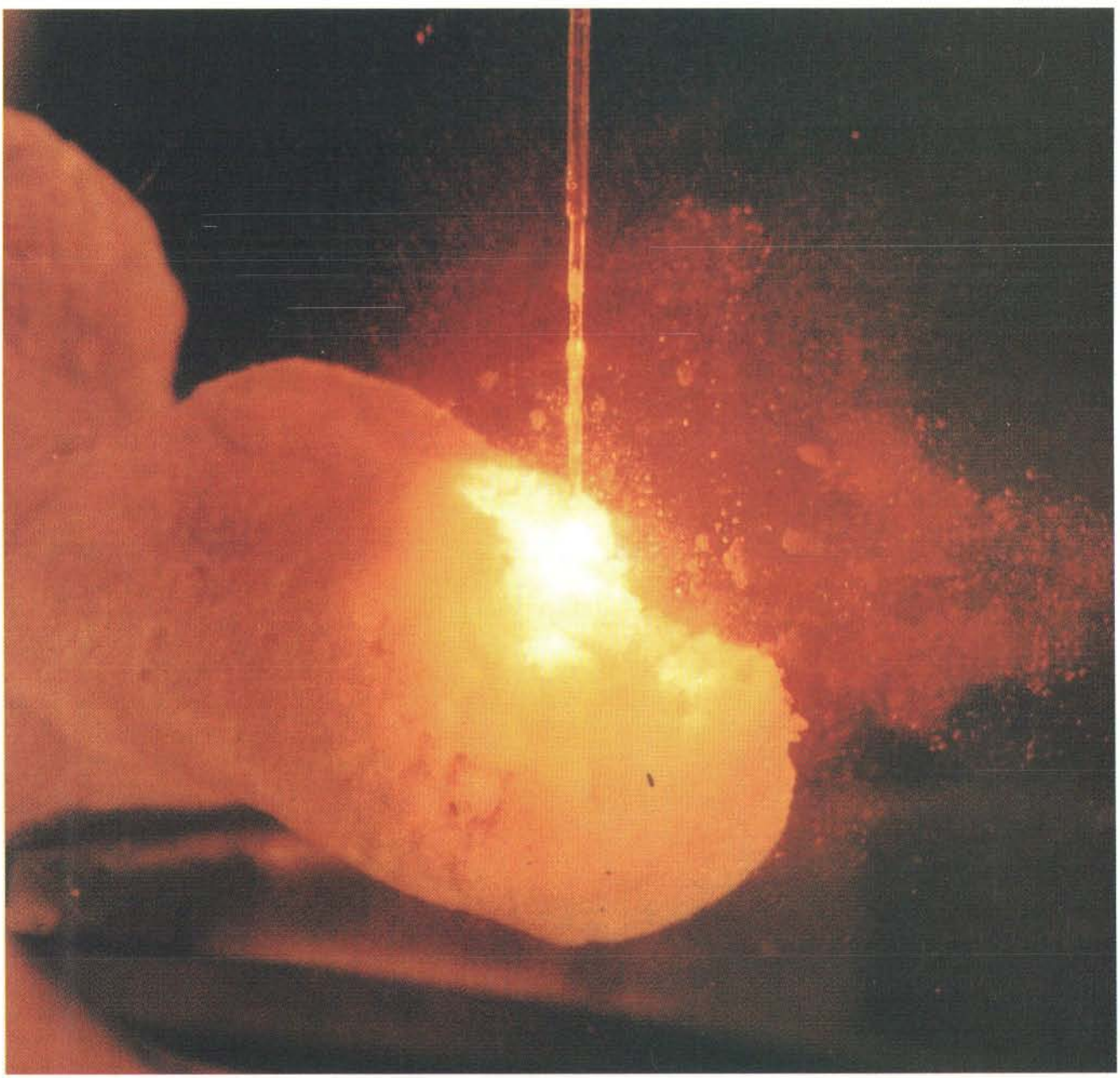
Yet in some circumstances, even the wider damage caused by the longer, slower heating of tissue can be turned to advantage. A general surgeon may

want to remove a damaged portion of the liver without causing extensive bleeding. Or a gynecologist may want to make a deep incision to excise an early-stage malignancy of the cervix and simultaneously use the heat to seal the surrounding capillaries that contribute to bleeding. In both cases, the long exposure to a continuous-wave laser (as opposed to a short-pulsed laser) reduces bleeding precisely because heat spreads to the capillaries nearby. A CO_2 laser with a wavelength of 10.6 microns may be used for these procedures because it is absorbed by the compound most common to tissue: water.

Although a continuous, and therefore thermal, beam may be needed for certain medical procedures, pulsed lasers can also remove tissue. My colleague J. Stuart Nelson has demonstrated that the erbium-yttrium-aluminum-garnet (YAG) laser, which has a wavelength of 2.9 microns and a pulse duration of 200 microseconds, can cleanly ablate calcified bone. On the other side of the visible spectrum is the xenon chloride excimer laser, which is in the ultraviolet region of the spectrum at 0.308 micron and has a pulse duration of 10 nanoseconds (billionths of a second). It can vaporize bone with little or no associated thermal damage.

Whereas these two lasers appear to affect tissue very similarly, they work in different ways. The energy of the ultraviolet photon is 10 times greater than that of the photon from the erbium YAG laser, yet the energy probably serves to break molecular bonds in the target through a nonthermal process known as molecular photodissociation.

When the tissue and its cells absorb the intense light of a laser, energy must be dissipated in some way. This release may take place in the form of heat,



URINARY TRACT STONE is among the many calcified deposits, including kidney stones and gallstones, that can be destroyed by lasers. Delivered in this instance through a 400-micron optical fiber, energy from the laser beam ultimately creates an ionized plasma. This plasma, in turn, produces shock waves, which break apart the stone.

Laser's Thermal and Nonthermal Effects

When the wavelength of a laser beam matches the target's absorption band, much of the light is retained by the target (*a*); the remaining light is scattered, reflected and in some cases transmitted. In contrast, photons can pass through organs or cellular structures with different absorption bands without damaging them (*b*). The laser beam is most commonly used to heat the target (*c*) and destroy it. It can also have nonthermal effects: energy from photons can break molecular bonds (*d*) and create an ionized plasma, thereby causing a shock wave (*e*) that breaks apart mineralized deposits. When dye is concentrated in the target, laser light can cause the target to fluoresce (*f*) for diagnostic purposes, or it can interact with the dye so as to destroy the target (*g*).

as we have discussed above, or in the form of photodissociation, shock waves, chemical reactions or fluorescence. Physicians use all these mechanisms either to alter or to study cells and tissues in very exact ways for the diagnosis and treatment of disease. Each of these varying effects also allows scientists to perform subcellular microsurgery.

Furthermore, by coupling lasers with other technologies such as fiber optics, one can achieve nonthermal, as well as thermal, results in previously inaccessible parts of the body [see "Optical Fibers in Medicine," by Abraham Katzir; SCIENTIFIC AMERICAN, May 1989]. By employing fiber optics and hollow tubes, for example, surgeons can administer laser light through the wall of the chest to treat two major lung disorders: spontaneous pneumothorax and severe emphysema. In the first condition, normally healthy individuals develop a spontaneous leak, or rupture, in one of the lungs. Lasers can be used to seal the rupture—as discovered by my colleagues Akio Wakabayashi and Matthew Brenner of the University of California at Irvine. The need for conventional surgery is therefore obviated.

The same procedure can help treat emphysema, which, in one form or another, afflicts more than 10 million Americans. A CO_2 laser—channeled through a hole in the chest wall—is applied to the fragile blisters, called bullae, that cover large areas of the lungs. The laser heat shrinks the blisters, sealing the leaks and reducing the risk of additional ruptures. To date, 11 of 12 severe emphysema patients, who were too sick to undergo conventional surgery, have shown improvement after such treatment.

In another application of this technology, cardiologists and radiologists can thread blood vessels with a single 400-micron fiber—or a flexible bundle of up to 400 50-micron fibers—until they reach a blockage in the peripheral or the coronary circulatory system. The optical fiber then transmits the laser light, destroying the blockages and restoring normal circulation.

Laser angioplasty, as this procedure is called, was originally performed with thermal laser beams as an adjunct to balloon angioplasty. The laser probe could be used to open a channel through a totally or partly occluded blood vessel. A balloon was then inserted and inflated, further dilating the blood vessel.

The thermal approach was not successful, however, in removing the calcified deposits that are common to atherosclerosis. My colleagues Jonathan M. Tobis and Walter L. Henry of Irvine found that it was not uncommon for the hard deposits to deflect the fibers, causing the bundle to perforate the blood vessel. Also, as seen in the case of the port-wine stains, heat can damage more than its target. In some cases, it can injure the surrounding normal vessel wall.

In contrast, the pulsed 0.308-micron excimer laser seemed ideal for laser angioplasty, with or without balloons. Excited pairs, or dimers, of gas atoms,

such as argon and fluorine or xenon and chlorine, release energy, generating the excimer laser beam. (Excimer is shorthand for excited dimer.) Flexible quartz optics efficiently transmit the 0.308-micron excimer wavelength. James S. Forrester, Frank Litvack and Warren S. Grundfest of Cedars of Sinai Medical Center in Los Angeles were the first to show that the excimer laser could restore coronary circulation.

To date, there have been more than 2,000 laser-assisted coronary angioplasties performed in the U.S., with a complication rate no greater than that of nonlaser angioplasty procedures. Currently it appears that the rate of reocclusion is at least comparable with that of balloon angioplasty. In the long run, however, I expect a procedure that breaks down and disintegrates the hard plaque will be more effective than merely stretching the vessel with a balloon. Either technique is an alternative to expensive and more risky coronary bypass surgery, which often requires long hospitalization.

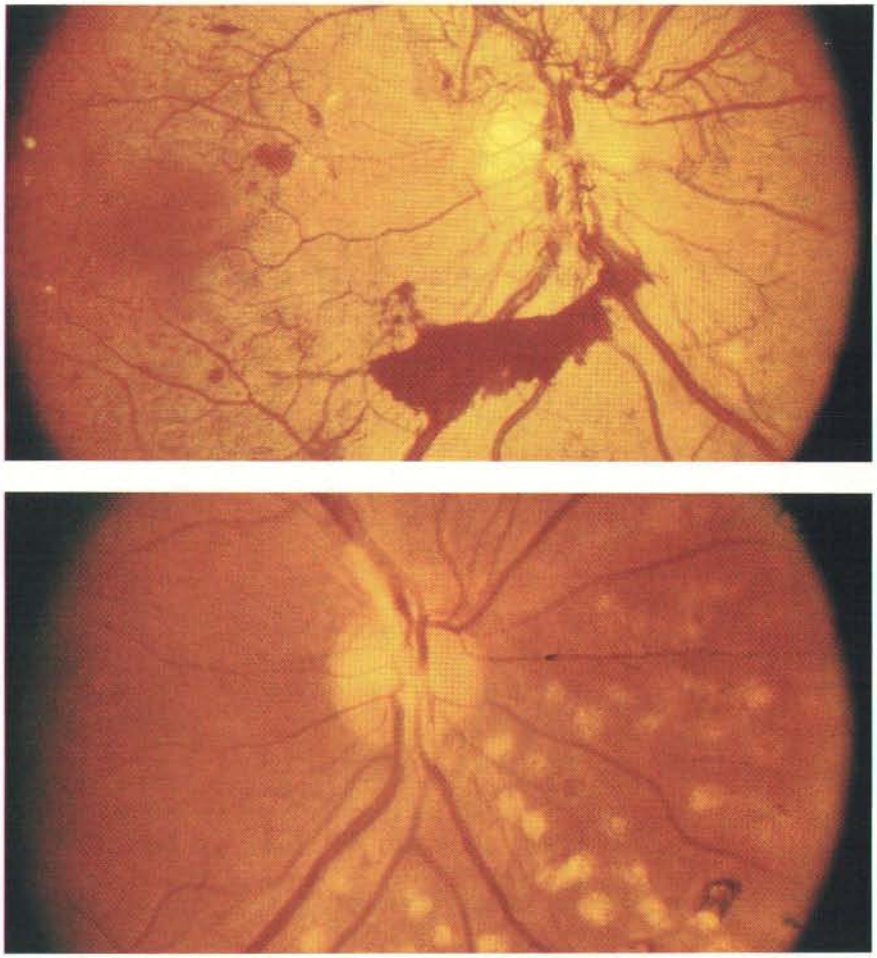
For now, laser coronary angioplasty remains experimental, and its widespread use must wait until the lasers and imaging equipment are entirely dependable. Eventually, digital imaging in combination with ultrasound technology may yield reliable methods of directly visualizing the surgery and guiding the laser.

Beyond the technological hurdles lie biological ones. The detrimental effects of the excimer laser are not entirely understood. For instance, the 0.308-micron wavelength can produce mutations. (These mutations may be more hazardous to the person operating the machine on a daily basis than to the patient, but adequate safety precautions should address this concern.)

Furthermore, no one is entirely sure how this laser ablates tissue. The energy of its photons may destroy by heating the target or by breaking molecular bonds (photodissociation). The individual photon energy of this wavelength is on the borderline between producing thermal and nonthermal effects.

In contrast, the reason for the argon fluoride excimer laser's destruction is clear: it breaks bonds. Photons from this laser, which has an ultraviolet wavelength of 0.193 micron, are roughly one third more energetic than those of the 0.308-micron wavelength. This energy is clearly enough to rupture individual molecular bonds. (The shorter the wavelength, the greater the energy associated with the photon.) Rangaswamy Srinivasan of IBM, who pioneered the use of the 0.193-micron ex-

PORT-WINE STAINS can be treated with lasers. Excess blood vessels just under the outer layer of the skin (*left*) absorb yellow laser light, which destroys the red vessels (*right*). Because the beam is delivered in brief pulses, other tissue is undamaged.



DIABETIC RETINOPATHY, which can lead to blindness, is characterized by excessive growth of blood vessels in the retina and some hemorrhaging (*top*). An argon laser can be focused onto the bloods vessels to stop the bleeding and then onto other areas of the retina where the laser light is absorbed by the brown melanin pigment. This procedure leaves circular burns (*bottom*), which by some poorly understood mechanism retard the growth of new blood vessels.

FLUORESCENT DYE is selectively retained in a mouse's skin, although it is visible only in the tail and ears. When the animal is illuminated by blue light, the dye luminesces. Similarly, dye that accumulates in cancerous tissue fluoresces when submitted to certain laser light, making the procedure effective for diagnosis.

cimer laser to etch thin polymer films, clearly demonstrated this ability.

The precision of photodissociation opened the way for laser surgery on the eye. In 1983 Stephen L. Trokel of Columbia University (in collaboration with Srinivasan) showed that by breaking bonds directly one could remove tissue in small increments from the cornea of the eye. Trokel and his colleagues illustrated the direct relation between the energy delivered by the laser and the removal of eye tissue: each pulse, 10 nanoseconds long, removed 0.2 micron of tissue (roughly one one-hundredth the diameter of a cell). The tissue removal was also strikingly clean and left the adjacent unexposed tissue apparently undamaged.

Trokel's remarkable findings led to the adaptation of the 0.193-micron excimer laser to corneal sculpting. The procedure is aimed at correcting major visual defects by reshaping the cornea. Surgeons can do this either by making small, deep linear cuts or by ablating shallower, wider areas. Eventually, if these procedures are successful, visual defects such as nearsightedness and farsightedness or astigmatism could be corrected. Eye surgeons also use the device to eliminate corneal scars and extra tissue growth. Each of these excimer laser procedures, regarded as experimental and controversial, is currently being tested in human trials.

Many questions need to be answered before such sculpting could become routine. Researchers remain unsure whether the surgery can cause mutations, whether the cornea can heal adequately, what the long-term effects of corneal damage may be and whether the equipment can be made precise and dependable. The cost of the laser systems—often as much as $500,000—is also high. Because many of the operations will be considered cosmetic by most insurers, patients may have to bear the cost.

In addition to breaking molecular bonds, lasers can induce shock waves, another medically promising nonthermal effect. Shock waves are useful for such ocular surgery as the removal of secondary cataracts, those that form on the membrane just behind the artificial lens in 30 percent of patients who have had such implants.

Prior to 1980 the only treatment for secondary cataracts was the surgical opening of the posterior membrane, a maneuver that requires general anesthesia. Then Danièle Aron-Rosa of the University of Paris and Franz Fankhauser of the University of Bern in Switzerland showed that short-pulsed infrared lasers could be focused on or near the opaque posterior membrane, causing it to be torn apart by a shock wave. They used the neodymium YAG laser, which can issue pulses on the order of nanoseconds or picoseconds (trillionths of a second) and operates at a wavelength of 1.06 microns.

After laser treatment, a patient's vision is almost always instantly improved. More than 200,000 such procedures, called posterior capsulotomies, are performed every year in the U.S. Unlike the earlier surgical procedure, which cost $2,000, the laser surgery costs less than $1,000 and requires no general anesthesia or hospitalization.

This application of the short-pulsed laser is possible because of the intensity of the narrowly focused beam: millijoules of energy are delivered for a period that ranges from 10^{-12} to 10^{-9} second on a spot 25 to 50 microns in diameter. The beam passes through the outer cornea and the artificial lens, which are both transparent to the 1.06-micron wavelength, and focuses selectively on the vitreous, or gel-like, substance next to the secondary cataract. (In certain laser procedures, the intensity of the beam may reach a level high enough to produce damage only at the point where it is focused. Thus, the light can pass harmlessly through overlying tissue or structures.)

In the case of posterior capsulotomy, the photons pass through the cornea and lens. At the focal point near the secondary cataract, however, the photon density is so high that electrons are stripped from atoms in a process called optical breakdown, or ionization. The electrons form a highly excited gaseous cloud, or a plasma, that captures the remaining photons from the laser. As a result of this absorption, the temperature in the focal spot quickly rises to tens of thousands of degrees Fahrenheit. The plasma then rapidly expands, causing a shock wave to propagate in all directions. The shock wave disrupts the secondary cataract. The plasma, however, is not created until the photons reach a certain threshold of intensity. Once that happens, the strength of the shock wave becomes proportional to the amount of energy absorbed.

Stones in the kidney, ureter and gallbladder are also candidates for shock-wave therapy. Surgeons delivering a short-pulsed beam through the urethra to the ureter by means of fiber-optic technology can break up hard deposits there. Gallstones can also be fragmented in a new, similar procedure. Surgeons insert an optical-fiber endoscope, called a laparoscope, through a small hole in the patient's abdomen. Using an electrocautery knife or a continuous-wave thermal laser beam, they can cut the diseased gallbladder out of the surrounding liver and pull it through the small incision.

Although lasers may not be necessary for the dissection, they can be very helpful if the gallbladder is laden with stones. These hard deposits often prevent surgeons from pulling the gallbladder through the small hole. In such cases, a pulsed laser is passed through the laparoscope directly into the gallbladder. The stones are then broken apart by shock waves, and the gallbladder can be easily removed. The operation is much less traumatic to patients than major surgery.

The immediate destruction of a cataract and fragmentation of a kidney stone are dramatic uses for lasers. But these devices can also achieve more subtle effects. Lasers can drive chemical reactions, just as the sun drives photosynthesis. In fact, laser photochemistry may be on the verge of becoming a viable treatment for cancer.

Again, this curative use of light is the recent incarnation of old observations. At the turn of the century, scientists noticed that cancerous tissue concentrated some of the body's own pigments, particularly porphyrin, the brownish-red pigment of blood. This observation was not exploited until the 1970s. Thomas J. Dougherty and his colleagues at Roswell Park Memorial Institute in Buffalo, N.Y., showed that animals injected with porphyrin retained high concentrations of the pigment in their tumors between 48 and 72 hours later. When light of a wavelength that matched the absorption band of the porphyrin was shined on the tumor, the growth darkened and often disappeared.

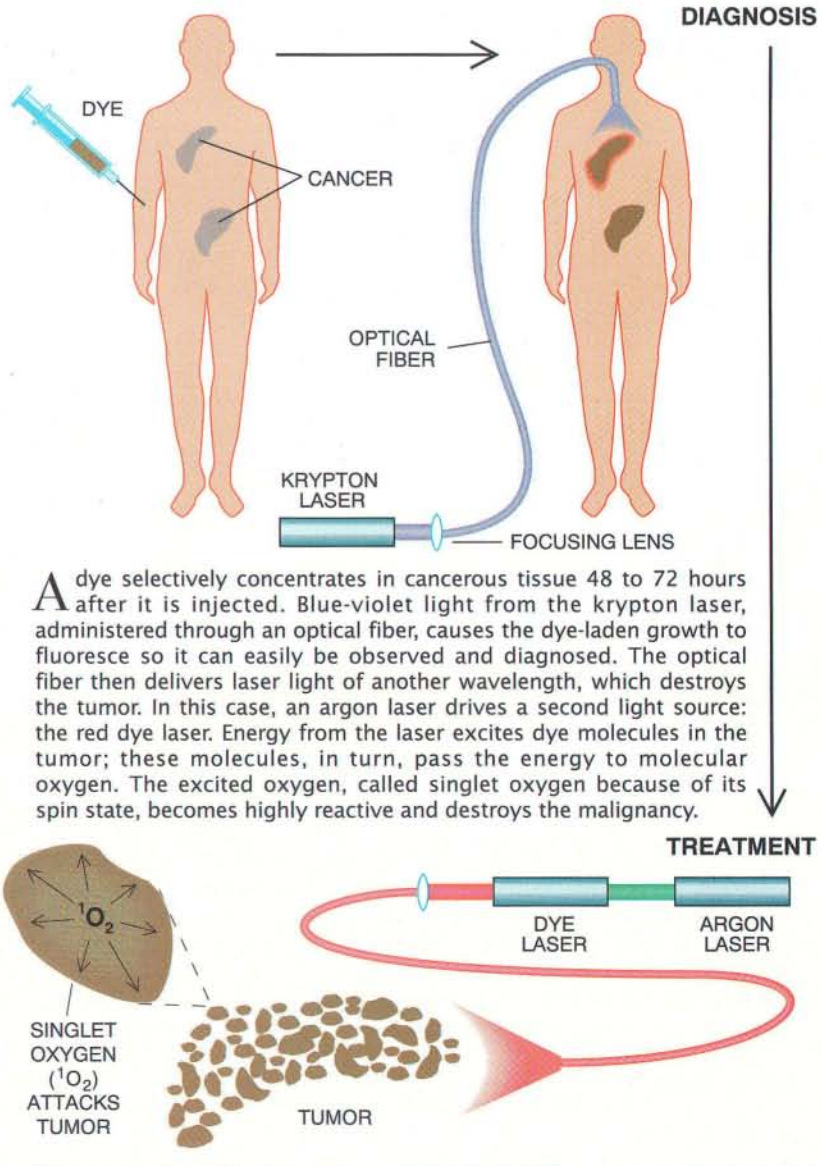
Researchers now know that the dye-mediated mechanism of tumor destruction works because of the generation of a cytotoxic, excited oxygen molecule, called singlet oxygen. ("Singlet" refers to the spin state of the oxygen molecule.) The transfer of energy from the excited porphyrin molecule to the oxygen creates the excited singlet oxygen. In this state, oxygen is highly reactive and therefore toxic. The singlet oxygen attacks the outer cellular membrane as well as many of the important intracellular, membrane-bound structures, such as mitochondria and lysosomes. Ultimately it causes the demise of the tumor without harming normal tissue.

Although early anecdotal reports were encouraging, the role dye sensitization, or photodynamic therapy, will play in cancer therapy remains to be determined. Long-term survival studies need to be completed. Comparisons need to be made with accepted treatments, such as radiation therapy, chemotherapy and surgery.

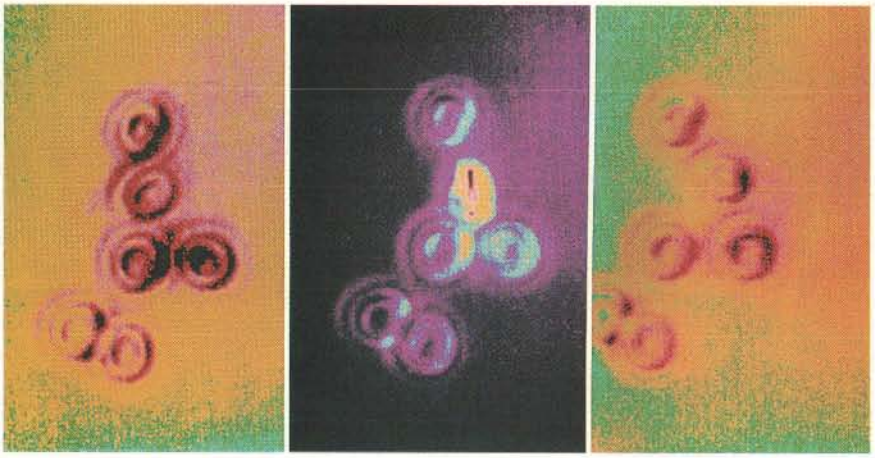
The U.S. Food and Drug Administration currently regards the use of photodynamic therapy for cancer to be experimental. One major drawback to the initial efforts was that patients became highly sensitive to bright light for as long as three months after dye injections. The development of better modes of application and new dyes that do not create residual photosensitivity are under way. Several major multicenter trials in the U.S. and Canada are exploring such treatment for lung, bladder and esophageal cancer.

The technique's diagnostic potential could prove more exciting. Many dyes that absorb energy can reemit the light as fluorescence, flagging cancerous areas. In preliminary studies in the U.S. and Japan, fluorescence has been used to detect early-stage lung and bladder cancer. By using fiber optics and sensitive light detectors, it may eventually

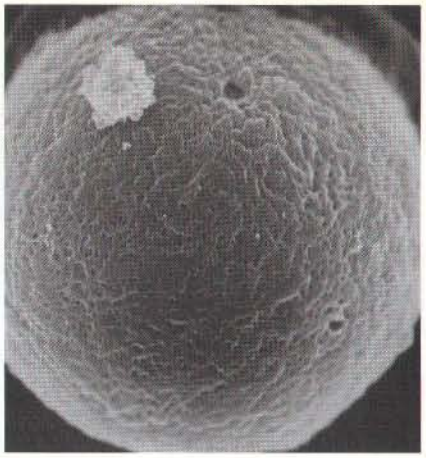
Photodynamic Therapy for Cancer



A dye selectively concentrates in cancerous tissue 48 to 72 hours after it is injected. Blue-violet light from the krypton laser, administered through an optical fiber, causes the dye-laden growth to fluoresce so it can easily be observed and diagnosed. The optical fiber then delivers laser light of another wavelength, which destroys the tumor. In this case, an argon laser drives a second light source: the red dye laser. Energy from the laser excites dye molecules in the tumor; these molecules, in turn, pass the energy to molecular oxygen. The excited oxygen, called singlet oxygen because of its spin state, becomes highly reactive and destroys the malignancy.



OPTICAL TWEEZERS are one of the many cellular uses for lasers. Photons from laser beams can hold cells immobile in an optical trap. The light can then push, pull or turn the tiny object. In this instance, the second red blood cell from the top is picked up, stood on its edge and flipped over.



RABBIT EGG outer layer, or zona pellucida, has been perforated twice by a laser to facilitate the entry of sperm for fertilization.

be possible to diagnose small cancers in internal regions of the body. Ultimately a patient may be given a "cocktail" of dyes, some for diagnosis and others for light-activated tumor destruction.

While the human body and its organs remain the primary focus of most laser research in medicine, smaller targets also engage workers' attention. In fact, lasers offer scientists a novel means of studying individual cells as well as the photophysical processes that provide the foundation of laser medicine. For 20 years, my colleagues and I have pursued such work [see "Cell Surgery by Laser," by Michael W. Berns and Donald E. Rounds; SCIENTIFIC AMERICAN, February 1970]. Understanding how light interacts with the cell and its organelles has contributed to the study of basic cell structure and function. Virtually all the photophysical mechanisms described in this article can be generated and studied at both the cellular and subcellular levels.

For example, photoactive dyes can be incorporated into a portion of the DNA molecule. This region of the DNA can then be inactivated by exposure to, say, the blue-green beam of the argon laser, resulting in a selective genetic deficiency. Alternatively, a highly focused, pulsed ultraviolet laser can generate a microplasma in the outer cell membrane, creating a small hole that will open up just long enough to allow DNA, coding for specific genes, to enter the cell. This DNA can be incorporated into the genetic material of the cell in a process 10,000 times more efficient than conventional genetic engineering. This technique may have its greatest value in the genetic engineering of crop plants, because it is extremely difficult to introduce foreign DNA across a rigid cell wall using only biochemical methods.

Researchers are also investigating a scaled-up version of cell perforation as a way to facilitate sperm penetration of eggs. Clinically it may be possible to use a highly focused, pulsed ultraviolet laser to drill a one- to 10-micron hole through the outer protective zone of a human egg. The sperm can then quickly enter. Preliminary studies conducted by Ricardo H. Asch at Irvine and Yona Tadir of Tel Aviv University have demonstrated the feasibility of laser micromanipulation of the egg.

In addition, Aaron Lewis, Neri Laufer and Daniel Palanker of the Hebrew University in Jerusalem were able to guide a 0.193-micron excimer laser beam to create holes in the outer layer of a rabbit egg. Fertilization took place at a higher rate than normal. The laser approach might be quicker and more selective than other methods of assisted sperm-egg penetration. But extensive studies must be done to examine the potentially harmful mutagenic and biochemical effects of laser energy.

Perhaps the most exciting new application of the laser in cell surgery and manipulation is its use as an optical tweezer. Arthur Ashkin of AT&T Bell Laboratories first described the ability of lasers to trap, or hold, an object in a beam of light. The force associated with the photons can be used to hold cells still, rotate them or pull them in any direction.

The ability to restrain cells or their internal organelles in an optical trap facilitates surgery with a second laser beam—the tiny object cannot slip away. Membranes of two adjacent cells can easily be fused together, or the tweezers can hold a chromosome while a second laser beam cuts off a fragment that could be isolated for cloning. In the case of egg fertilization described above, it may become possible to trap a sperm and move it to the newly created hole in the egg's outer layer.

Whether surgery occurs at the level of the organ or the cell, lasers offer physicians and researchers an unprecedented and precise set of optical tools. Indeed, in many specialties, laser surgery has already become a standard procedure that medical residents learn to use. Only time and good clinical studies will tell how these applications can be expanded.

FURTHER READING

SELECTIVE PHOTOTHERMOLYSIS: PRECISE MICROSURGERY BY SELECTIVE ABSORPTION OF PULSED RADIATION. R. Rox Anderson and John A. Parrish in *Science*, Vol. 220, pages 524–527; April 29, 1983.

EVALUATION AND INSTALLATION OF SURGICAL LASER SYSTEMS. David B. Apfelberg. Springer-Verlag, 1986.

INTRODUCTION TO LASER PHYSICS AND LASER TISSUE INTERACTIONS. J. Stuart Nelson, William H. Wright, John Eugene and Michael W. Berns in *Endovascular Surgery*. Edited by Wesley S. Moore and Samuel S. Ahn. Harcourt Brace Jovanovich, 1989.

IEEE JOURNAL OF QUANTUM ELECTRONICS: SPECIAL ISSUE ON LASERS IN BIOLOGY AND MEDICINE, Vol. 26, No. 12; December 1990.

LASER MICROBEAM AS A TOOL IN CELL BIOLOGY. Michael W. Berns, William H. Wright and Rosemarie Weigand Steubing in *International Review of Cytology*, Vol. 129, pages 1–44; 1991.

Music listeners can hear dramatic 3-dimensional sound from conventional mono and stereo recording or broadcast sources, thanks to a sound reproduction technique developed by Hughes Aircraft Company. This Sound Retrieval System (SRS) creates the ambiance and dynamic range of a live performance or studio recording. It retrieves and restores spatial information using real-time processing techniques that, like the human ear, recognize the direction from which a sound originates. Because its circuitry has been reduced to a single microchip, SRS is likely to be incorporated into a wide variety of audio products.

In a major breakthrough in integrated circuit technology, Hughes has developed a technique for producing distinct lines approximately one two-millionth of an inch on semiconductor chips. These ultrasmall features, which are 100 times smaller than most commercial integrated circuits, will play a vital role in an emerging integrated circuit technology based on quantum physics. Rather than using electron beams, they were created with a focused ion beam, since features in resist material can be defined much more accurately using ions. Scientists predict these semiconductor chips will operate 10 times faster than conventional circuits.

Pilots flying special operations helicopters on low-level missions in total darkness, smoke and fog, will be aided by the field-proven Hughes Night Vision System, designated the AN/AAQ-16. HNVS is being installed on U.S. Army MH-47E Chinooks and MH-60K Blackhawks, on U.S. Air Force MH-60G Pavehawks, and a derivative of the system has been selected for the Marine Corps' V-22 tilt rotor aircraft. The system, produced by Hughes, has been installed on several other military helicopters, including the U.S. Navy's SH-2F Light Airborne Multi-Purpose System (LAMPS) MKI. The turret mounted infrared system provides the crew with TV-like imagery on a cockpit panel display.

A rescued communications satellite is seeing space service once again. Westar VI, recovered in 1984 by American astronauts, was refurbished by Hughes to serve new markets in Asia. The satellite was restored to flight condition for Asia Satellite Telecommunications Co., Ltd. (AsiaSat) and renamed AsiaSat I. It is providing domestic telecommunications for China, Thailand, Pakistan, and other Asian countries. Hughes has also refurbished another recovered satellite, Palapa-B2R, for use by Indonesia. Both refurbished spacecraft were successfully launched in April, 1990.

Hughes Aircraft Company's Missile Systems Group has excellent opportunities for Electronics Engineers. We're a world leader in developing and manufacturing advanced tactical missile systems, airborne avionics, launchers, weapon control systems, guidance and propulsion systems, and field support and test equipment. Applicants should have a background in Computer Science or Physics, including 3-5 years experience with an emphasis in simulation and analysis. Please send resume to: Hughes Aircraft Company, Missile Systems Group, Attn.: Employment Dept., 8433 Fallbrook Avenue, Canoga Park, CA 91304-0445. Proof of U.S. citizenship may be required. Equal opportunity employer.

For more information write to: P.O. Box 45068, Los Angeles, CA 90045-0068

HUGHES

Subsidiary of GM Hughes Electronics

Arthur Stanley Eddington

He led an astronomical expedition to test Einstein's theory of general relativity, expounded the idea of an expanding universe and inferred the internal structure of the stars

by Sir William McCrea

It is hard to convey to the present-day reader the widespread respect that was given to Arthur Stanley Eddington in the years between the world wars. He exerted an enormous influence on the development of physical thought. The influence came first from his own contributions to astronomy and astrophysics and second from his insights into the contributions of others; he often seemed to grasp the significance of advances more profoundly than those who made them and to explain the advances more skillfully. Further, his attempts to expose the foundations of physics have continued to influence the thinking of scientists of subsequent generations, most of whom never realize that they may be pursuing a quest started by Eddington.

In the 1920s and 1930s outstanding personalities dominated every field: Churchill and Roosevelt in statecraft; Shaw and Eliot in letters; and Bohr, Einstein, Rutherford and Eddington himself in physical science, to name a few. There are no such giants today. One is not saying that comparable talents do not exist, only that, for reasons that could be enlarged on elsewhere, such talents would now be much less likely to win so much esteem. Certainly in science, the pursuit of knowledge has devolved from the individual to the team and, to some extent, from the originators of ideas to the presenters of ideas. In all honesty one must admit that Eddington, perhaps more than anyone else, pioneered the role of presenter. But he was in the first place an originator, and it is as such that one thinks of him here.

Bohr, Einstein and Rutherford had made great discoveries, but they played only relatively minor roles in expounding their work outside professional circles. Eddington was different. Quite a few scientists heard him lecture; many more read something by him; and most of them were convinced that they had understood him. Eddington had the aspect of one who enjoyed special access to the secrets of the universe, and what he set down came to be regarded by many as verging on the status of holy writ.

Of course, the lasting value of his work does not depend on the almost mystical air he bore at the height of his powers. But his charismatic appeal did attract attention to his ideas. That appeal seems to have come from the quite enormous faith he put in his own intuition, in part because it had so often turned out to be justified.

Eddington was born to Quaker parents in 1882, lost his father two years later and studied mathematics and physics at Owens College (now Manchester University) and then at the University of Cambridge, where he was Senior Wrangler—the highest mathematical honor. He worked at the Royal Observatory at Greenwich from 1906 to 1913, when he became Plumian Professor of Astronomy and Experimental Philosophy at Cambridge. Six years later he led a famous expedition to an island near equatorial Africa to observe a solar eclipse that would test Einstein's new theory of general relativity. He served as president of the Royal Astronomical Society in London, was knighted in 1930 and died in 1944.

Until his last illness, Eddington was physically vigorous. He enjoyed swimming and playing golf (not expertly, one gathers). A noted cyclist, he kept a careful record of his longer rides, from which he evaluated his parameter n, the number of days on which he had cycled n miles or more. I think the value finally reached 75.

Eddington was painfully shy in casual social encounters and poor at extemporaneous speaking and classroom lecturing. But a prepared public lecture he delivered with sparkling wit, and he was masterly in presenting a concise summary of a research paper at a scientific meeting. Those whom he knew well considered him a clubbable and caring companion. When, for instance, his lifelong colleague, the astronomer F.J.M. Stratton, had a serious illness in middle life that threatened to leave him depressed, Eddington almost daily invented an intellectual conundrum to keep Stratton alert and so to speed his recovery.

Quakerism helped to shape Eddington's outlook on life, and it might appear almost inevitable that it should remain congenial to him. But things often go the other way: many men of independent mind react against their upbringing. So clearly it was by deliberate decision that Eddington maintained this attitude, and he discussed it quite freely and fully in a number of his lectures and addresses, emphasizing the essential element of "seeking" in the Quaker tradition.

I think it correct to say that Eddington never suggested that his study of physics and the physical universe in fact had much influence on what one might call his religious views. Doubtless he would agree that all knowledge is one and what we seek to discover is how all things are connected, but he also asserted that scientists may have a long way to go before they can see the essential connections. "In science as in

SIR WILLIAM McCREA is professor emeritus of theoretical astronomy at Sussex University, a former president of the Royal Astronomical Society (a post Eddington had held) and a fellow of the Royal Society. He was knighted for his scientific achievements in 1985. "In 1923," he says, "when I was a freshman at Cambridge, a friend pointed out Eddington as the man who knew more about relativity than anyone but Einstein. In 1926 the astrophysicist E. A. Milne advised me, 'Whatever else you do, read Eddington's *Internal Constitution of the Stars*,' which was about to be published. Although for some while I stayed in mathematical physics, the book influenced me in my decision to switch to astrophysics, and thus led, indirectly, to the writing of this article."

religion the truth shines ahead as a beacon showing us the path; we do not ask to attain it; it is better far that we be permitted to seek," as he said in his Swarthmore Lecture in 1929.

Eddington's convictions made him a conscientious objector during World War I. The Astronomer Royal, Sir Frank Dyson, was very eager that Eddington might be available to observe the solar eclipse of 1919, should world conditions permit. To this end, he used his influence to have Eddington kept at his post at Cambridge through the war. A number of other prominent scientists held views similar to Eddington's during World War I. Most of these, including Eddington, felt bound to take a different stance in World War II.

Eddington seems to have had no settled scientific objective before going to the Royal Observatory in 1906. But there he joined with enthusiasm in the observatory's work, initiating, for example, the program for the determination of latitude variation, which in one form or another has gone on almost to the present.

For most of the next 10 years, Eddington devoted himself to analyzing stellar motions, a study that had been initiated in 1904 by the astronomer J. C. Kapteyn, of the Netherlands. Eddington reviewed the entire subject, including his own major contribution, in his first book, *Stellar Movements and the Structure of the Universe,* published in 1914. It is interesting to see what the universe was taken to mean then: it was the "stellar system," or "galaxy," envisaged as a flattened aggregate of stars in which the sun occupied "a fairly central position." The sun was encircled by the Milky Way, consisting of stars, nebulas and clouds of obscuring matter.

The system as a whole appeared to be in a steady state, a picture that induced Einstein to model the cosmos as a static entity by introducing the so-called cosmological constant into the original version of general relativity. Eddington was to seize on this constant—which Einstein later renounced—as indispensable to all his subsequent work.

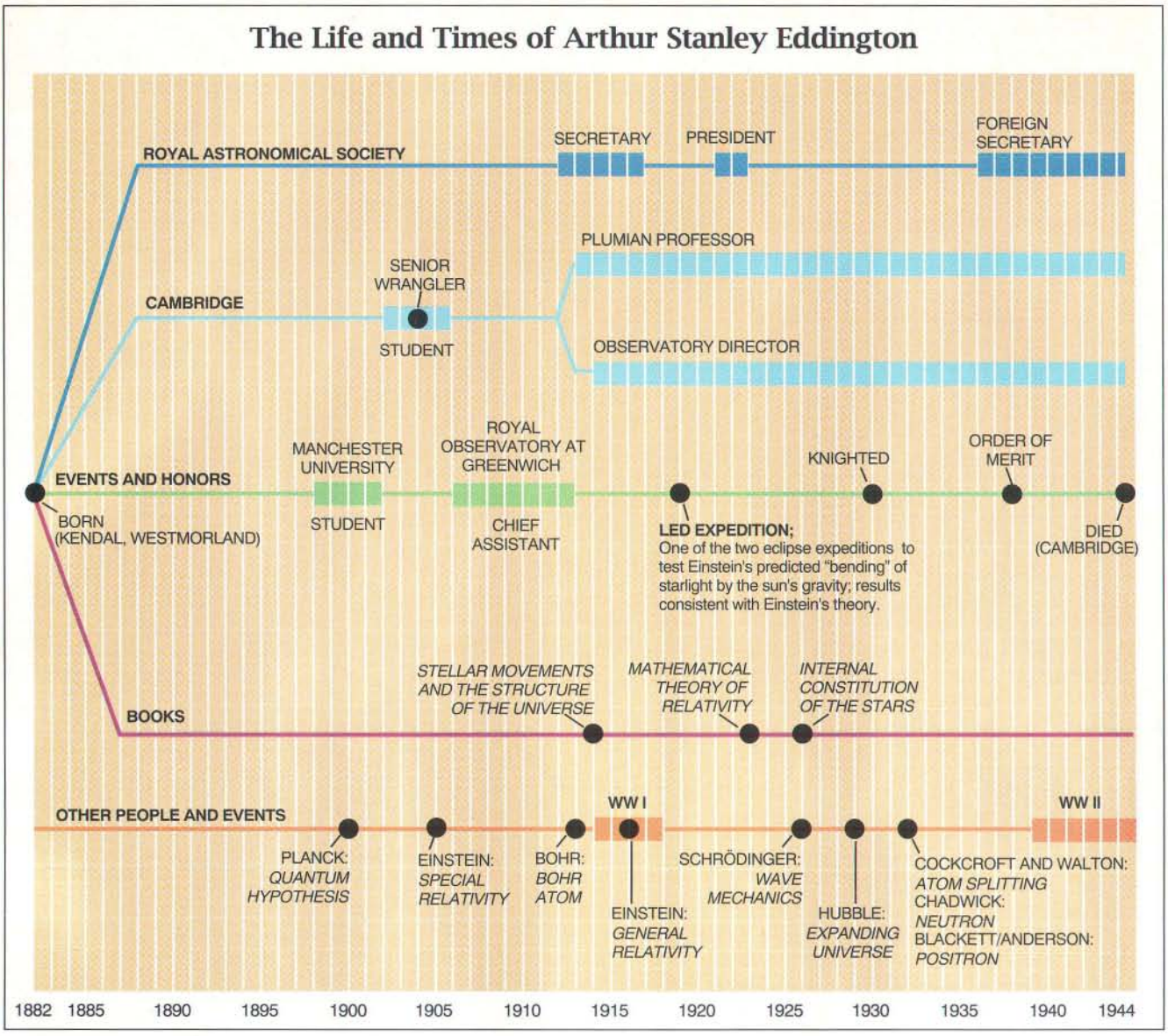
The main part of the book, on stellar kinematics, provided the starting point for voluminous investigations by subsequent workers. The last chapter proved to be the stimulus for other vast work on stellar dynamics. Eddington also gave what was probably the first properly informed discussion of what were then called spiral nebulas. He favored—correctly, of course—the view that they are galaxies like our own. Finally, he called attention to the necessity of supposing, as Henri Poincaré had already suggested, the Milky Way to be rotating. In a word, Eddington was the first to systematize what was then known about the large-scale structure of the observed universe.

EDDINGTON was at the height of his powers in the early 1920s, when he sat for this presidential photograph for the Royal Astronomical Society in London.

Eddington had his first brush with nascent general relativity in 1912, three years before Einstein was to give his full account of the theory. Einstein had produced certain interim results, one of which predicted that starlight passing near the sun would be deflected by the solar gravitational field. Such a deflection, Einstein noted, should be observable when a total eclipse obscured the solar disk, that is, if the starlight proved bright enough to be seen through the sun's corona. The displacement Einstein gave at that stage in his research was in fact only half the value ultimately predicted by general relativity, and indeed, it could have been derived from a slight adaptation of Newton's theory of gravitation.

In 1912 Einstein's correspondent and future colleague E. F. Freundlich started making preparations to test Einstein's prediction at an eclipse that would be observable in eastern Europe in 1914. He applied for technical assistance to the joint eclipse committee of the Royal Society and the Royal Astronomical Society, of which Dyson was the chairman and Eddington was a member. The committee regretted having to tell Freundlich that it did not have available the special equipment which he sought. But Dyson and Eddington had been alerted as to what was afoot.

The war in Europe started before Freundlich could observe the eclipse; it also cut off normal flow of scientific literature from Germany to England. When Einstein's great paper on general



relativity appeared in 1915, only one copy reached England. It had been sent by the eminent Dutch astronomer W. de Sitter in the neutral Netherlands to Eddington, who was then secretary of the Royal Astronomical Society. De Sitter followed the paper up with three fundamentally important papers of his own, in which he expounded Einstein's latest ideas and certain developments by himself. These papers were the first on general relativity to be published outside Germany.

The next was Eddington's *Report on the Relativity Theory of Gravitation,* submitted to the Physical Society of London in 1918. This work showed Eddington's astonishing facility for assimilating and presenting entirely new and revolutionary concepts and mathematical techniques. This report served as the basis for his famous *Mathematical Theory of Relativity,* published in 1923. More people must surely have learned general relativity through that book—either by reading it themselves or learning it from someone who had learned it from the book—than in any other way.

Dyson pointed out that the solar eclipse of May 1919 would occur against what would probably be the most favorable background of stars possible, a rich patch of the bright Hyades. Fortunately, the ending of the war made it possible for Dyson to organize two British expeditions to carry out the observations. Eddington led one of them to the island of Príncipe, off the west coast of Africa, which was the first to report its results. The other expedition went to Sobral in northeastern Brazil. Together, the findings were generally accepted as being in convincing agreement with Einstein's prediction because they were close to his calculations and unambiguously different from those based on Newton's theory.

The measured effects were displacements on photographic plates by a few hundredths of a millimeter. Since then, some physicists have asserted that Eddington's techniques had margins of error too large to prove Einstein right and Newton wrong. This is unfair: confirmation is not proof, but it is all that experimentalists have to offer. More sensitive techniques had to await the development of radio astronomy after World War II.

The vindication of a modification of Newton's law of universal gravitation caused worldwide excitement—possibly more than any scientific discovery before or since. Postwar internationalist sentiment contributed to the public acclaim: Britons were supporting the theory of a German-born physicist. Einstein leapt to instant fame. And if New-

ton's law had to be somewhat altered, it was recognized to be most fitting that the work of British astronomers should be responsible for altering it.

Eddington was the first person to see into the depths of a star—so one dares to claim. Geologic evidence shows that the nearest star, the sun, has been radiating energy at a quite steady rate for billions of years. What is going on inside the sun to enable it to do this? Eddington's contemporaries had been apt to argue, "You cannot know what it is like inside the sun until you know where it gets its energy, but you cannot know where it gets its energy until you know what it is like inside." Even the simple fact that the mean density of the sun is only about 1.5 times that of water was puzzling. What state of ordinary matter could allow such a massive body to have mean density no greater than that? And how does energy get through matter at that density? Most had supposed it to be transported by convection currents.

Eddington began his work in stellar structure in 1916, when he was trying to understand how the stars known as "Cepheid variables" produce their periodic variations in luminosity. He hypothesized that Cepheids function as spheres of gas that expand and contract. Following a suggestion made by R. A. Sampson in 1894, Eddington treated the transport of energy as radiative, not convective. This work was classic in quality. Moreover, its evident success suggested to Eddington that the gaseous sphere might be a feasible stellar model for more general purposes.

So Eddington set himself the problem of calculating the equilibrium state of a body of material of given mass and composition that behaved as an ideal gas and was held together by gravity, with energy transported by radiation. He assumed a formula for the dependence of the material's opacity on its nature and physical state and for the distribution through it of sources of energy generation. Now, the Swiss physicist J. R. Emden had calculated the equilibrium states of particular gravitating gas spheres known as polytropes. Eddington noticed that under certain plausible simplifications, his equations could be made to mimic Emden's, so that he could use Emden's numerical results. In this way, he obtained what he called his standard stellar model.

Eddington's model predicted a simple relation between the mass, radius and luminosity of a star, in which the effect of the radius was quite small. He found that the resulting mass-luminosity "law" fitted the observations for all the normal stars, including the sun, for which he had figures. Such stars were therefore seen to behave essentially like a perfect gas.

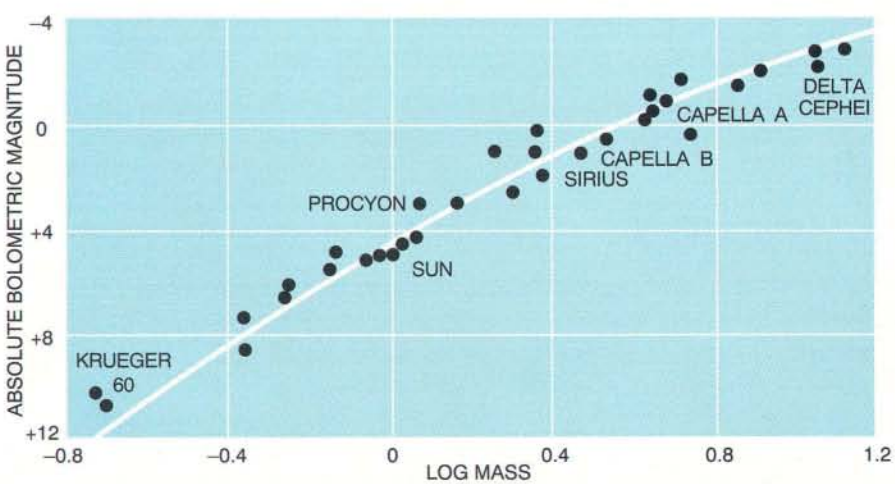
That material many times as dense as everyday liquids and solids, like that inferred to exist in central regions of stars, should obey the same gas law as the air we breathe was a revolutionary thought. This was Eddington's discovery. His friend J. H. Jeans had been the first to call attention to the fact that matter in stars must be highly ionized. Eddington saw how ionization enabled matter to behave like a perfect gas even at such high densities. Apparently he could then assert that a quantity of ordinary stellar matter equal in mass to that of, say, the sun will form a star just as bright as the sun. Thus, he seemed able to predict the energy flux without knowing its source.

Eddington explained his finding by arguing that radiative energy has to make its way out of the star against the opacity of its material. The opacity depends on the material's temperature and density according to a formula derived from simple quantum theory. Eddington postulated that the energy is generated in the material at a rate that depends on temperature and density. His calculations seemed to show that temperature and density must adjust themselves to give the total output expressed by his mass-luminosity law. If they cannot do this, the star will not be able to maintain a steady state, that is, we shall not have a star.

Eddington's model also implied that energy generated deep inside a star such as the sun takes, on average, about 10 million years to come to the surface and escape. Thus, were all energy sources in the sun completely switched off beginning today, it would take about 10 million years before anybody began to notice much difference. This delay between the cause and its ultimate visible effects was one reason Eddington could say so much about the interior of a star without knowing how the star gets its energy.

When Eddington presented his conclusions to the Royal Astronomical Society, they provoked lively controversy. Some astrophysicists followed Jeans in arguing that the light emitted by a star depends entirely on the sources put into it; if you do not know the one, you cannot claim to predict the other. Also, some still said, it is absurd to claim that material as dense as, say, a brick may be treated as a gas. A little later E. A. Milne, another great pioneer in astrophysics, argued that certain of Eddington's conclusions were consequences of restrictions he had introduced into his mathematics, rather than the physics of his problems.

The resulting debates at monthly meetings of the Royal Astronomical Society during the mid-1920s attracted unprecedented interest. Many of the country's most famous scientists attended—at least one leading mathematician became a Fellow of the Society just to have the right to be there. Eddington and Jeans gave each other no quarter; they could behave in this way—and enjoy doing so—because privately they were on excellent terms. When somewhat younger astrophysicists entered the fray, they often took things much more to heart. The immediate outcome as a whole was decidedly in Eddington's favor.



MASS-LUMINOSITY LAW, discovered by Eddington, relates the brightness of normal stars, such as the sun (*middle*) mainly to their masses (with only weak dependence on stellar radius). Eddington was able to make the prediction without knowing the means by which the stars generated their energy.

Eddington gave a comprehensive account of his views in *The Internal Constitution of the Stars,* published in 1926. It has proved to be a great classic of astrophysical theory.

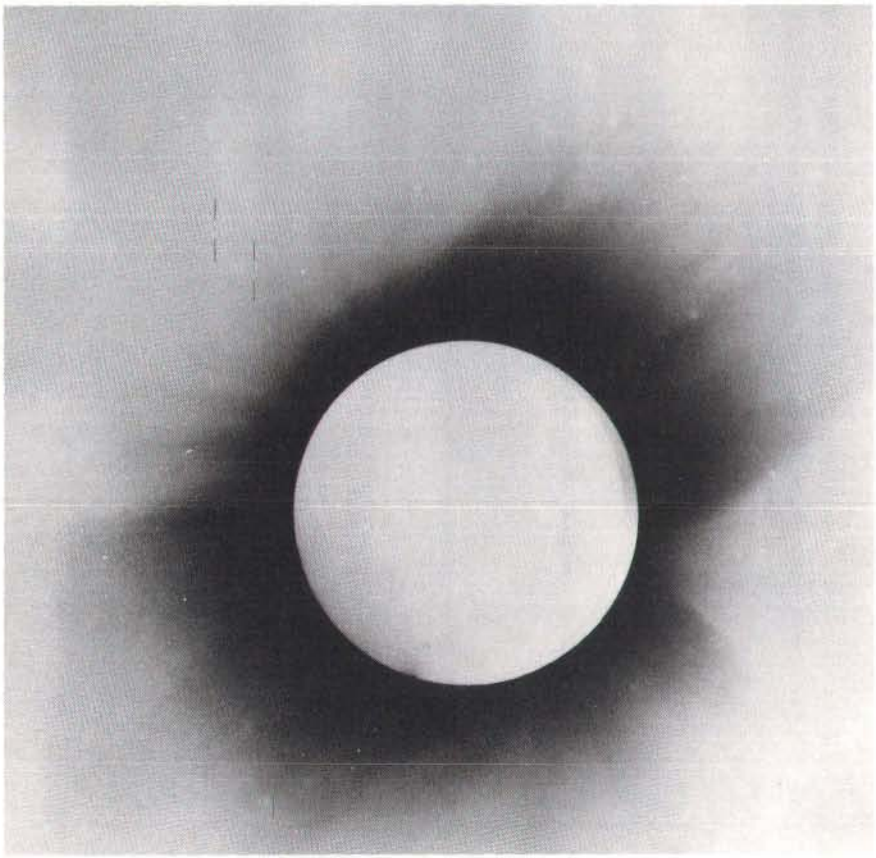
Eddington had been luckier than he knew. A dozen years later thermal nucleosynthesis became accepted as the process by which energy is generated in stars. This process is extremely sensitive to temperature. It can be shown that this feature in actual stars explains why Eddington was able to discover so much about their constitution before he knew much about the energy-generation process.

Most astrophysicists also believe that around 1935 Eddington went astray over one particular development. He himself had shown how his treatment encountered a fundamental impasse when he sought to apply it to the excessively dense stars known as white dwarfs, of which the most famous example was the faint companion of Sirius, itself the brightest star as seen from the earth, the sun excepted. The mathematical physicists at Cambridge immediately pointed out that the free electrons in so dense a star would form a degenerate gas in the quantum mechanical sense—that is, they would greatly resist further compression. The material would therefore obey an "equation of state" different from that used by Eddington for matter behaving "classically." R. H. Fowler showed that the difficulty is thus apparently completely resolved. Eddington welcomed this explanation.

Other theorists, however, went on to point out that the compression that produces degenerate electrons also raises their effective velocities nearly to the speed of light, so that relativistic effects soon become important. They showed that the equation of state for a relativistic degenerate electron gas is different again. Subrahmanyan Chandrasekhar, then at the University of Cambridge, made a remarkable discovery. A body of such matter cannot find an equilibrium state if its mass exceeds the "Chandrasekhar limit" of approximately 1.5 solar masses. Unless some process supervenes to break up such a body, it must undergo catastrophic collapse toward the state known as a black hole because its gravitational attraction prevents even light escaping from it. No known observational evidence conflicts with Chandrasekhar's conclusion.

Eddington would not accept the conclusion; he regarded it as a reductio ad absurdum of the theory. Most astrophysicists think that he was plain wrong; indeed, he identified no error in Chandrasekhar's work. So far as I know, however, even after more than 50 years no one has yet discussed precisely how a mass of real matter greater than the Chandrasekhar limit does actually evolve. Nor have any of the various candidates for black hole status, in observed stellar systems, been convincingly confirmed as such. We may here be working near the limits of applicability of standard relativity and quantum theory. If these have not been overstepped, presumably Eddington was indeed wrong. If Eddington was not wrong, then the others concerned must have somewhere passed the limits.

When Eddington embarked on his study of stellar constitution more than 70 years ago, knowledge of the subject was effectively nil. It is amazing that the picture at which he so soon arrived remains valid in all broad essentials to this day. Maybe he had a dash of luck and maybe he was mistaken about the consequences of relativisitic degeneracy, but the power and scope of Eddington's physical insight are still nothing short of awesome.



SOLAR ECLIPSE of 1919 confirmed Einstein's prediction that starlight passing near the sun would be deflected by the sun's gravitational field. The deflection amounted to only a few hundredths of a millimeter on this photograph (*negative*), taken at Sobral in Brazil. (The stars are marked by vertical lines.) Eddington led the other eclipse expedition to Príncipe, off the west coast of Africa.

In his first major astronomical study, which culminated in his book *Stellar Movements,* Eddington had elucidated the constitution of a galaxy, in effect of all galaxies. Then, in the *Internal Constitution,* he elucidated the constitution of the stars. And as stars form the building blocks of any galaxy, everything in modern astronomy must depend significantly on Eddington's discoveries. This is an enormous claim for the work of one man.

Yet Eddington went even further, for he showed that Einstein's original model of a static universe had to be unstable. He thereby established, according to general relativity, that the universe must be expanding (or contracting, but not static). He thus made an essential contribution to the astronomy of the cosmos in the large.

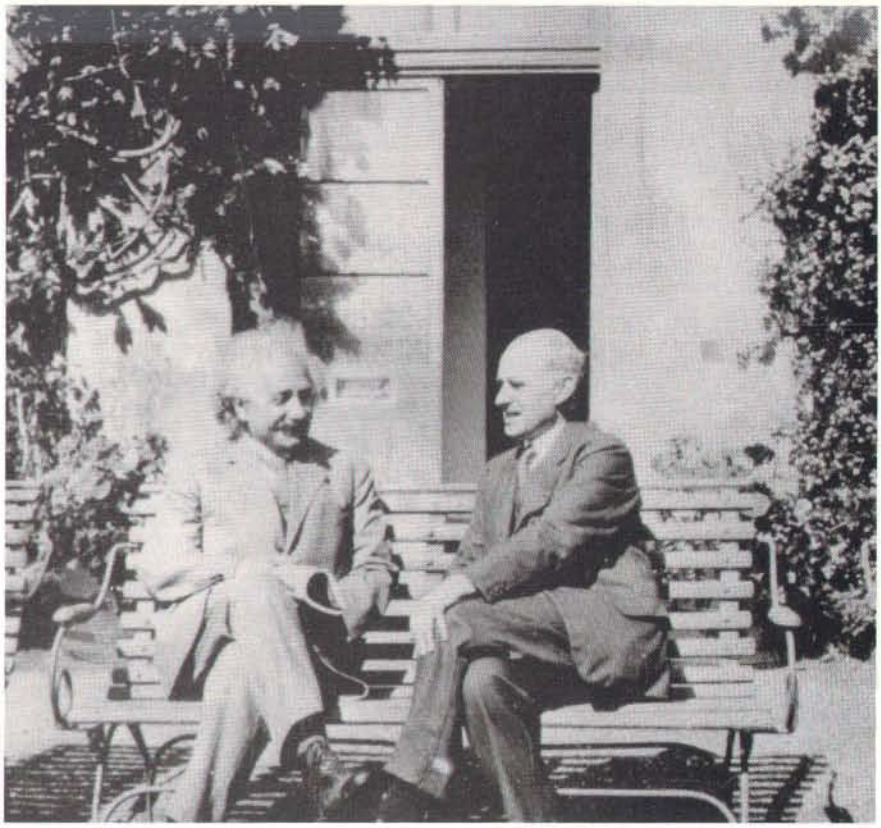
Eddington looked on his work in cosmology from a standpoint outside of cosmology. Indeed, he was coming to attach greater significance to the philosophical speculations he was then beginning to formulate than to all that he had achieved in astronomy as such. What is the universe? What is physics?

These were the fundamental questions to which Eddington devoted the rest of his scientific life.

In retrospect, we can see intimations of Eddington's metaphysical interest as early as 1928, when he wrote *The Nature of the Physical World.* His *Philosophy of Physical Science,* published in 1939, provided his more mature views in this realm.

He noted that certain numbers in physics do not depend on any system of units. Such "pure" numbers include the ratio of the electrostatic force between a proton and an electron to the gravitational force between them—a number of about 10^{39}—and the ratio of the mass of the proton to that of the electron, about 1,836. Standard physical theory makes the assumption that such ratios are universal constants, that "exact" values exist for them—whatever that may mean—and that these values can be approached but never indubitably established by more and more refined measurement.

Eddington questioned this point of view. He was convinced that there must be mathematical reasons for the numerical constants of physics to possess their particular values, and he proceeded to present them in his fundamental theory. In fact, the values he produced were in good—some in unbelievably good—agreement with experimental evidence. The development involved much clever mathematics, some of it invented for the purpose by Eddington, and it should be said that no one ever detected a serious mathematical error in anything he wrote.



GREAT CONTEMPORARIES: Einstein regarded Eddington's books on relativity as the best in English. Here the two are shown sitting in Eddington's garden in 1930.

Yet it is a fact that no particular results of this part of Eddington's work have been accepted. Nor has any physicist ever claimed to be certain of the postulates from which Eddington started or to follow his reasoning all the way through to any of his main conclusions; always there appeared to be some gap—some infuriating gap—in the logic.

"But it was in the formulation of the great series of questions ... that he gave his best gift to his fellow-workers," wrote his one-time student George Temple. "These questions form ... a great series of signposts leading to a Promised Land which perhaps he never quite attained in this life."

Eddington was a great man, a great scientist—and a great opportunist. Every great man has to exploit his opportunities; if he did not, we would never hear about him and we could not call him great, or anything else. Some of the ground had been prepared for him, as in stellar movements by Kapteyn and in stellar constitution by Emden. More generally, atomic physics was just reaching the stage at which Eddington could use it to understand, say, ionization and opacity in stellar material and even something about energy generation in a star. These examples illustrate the opportunities that Eddington's astonishing intuition led him to exploit with such fecundity.

In spite of Eddington's many common interests with his contemporaries at Cambridge, he could not be said to belong to any school of thought. Nor did he found a school of his own, although he certainly had a succession of highly distinguished and devoted graduate students. By the same token, of his more than 110 research papers, he had a collaborator in only seven.

Finally, even in his theory of the foundations of physics Eddington was one of a small cluster of workers independently pursuing essentially similar ends. Eddington, Einstein, Milne, Erwin Schrödinger a little later and Hermann Weyl somewhat earlier tried to construct unified theories of physics and cosmology and ultimately failed. One might say they wasted time and talent in such endeavors. Yet if one thinks of these men individually, it is difficult to see what else they might have done had they not turned to these challenges. Moreover, what they were trying to do is what some leading workers have been trying to do ever since—to seek grand unified theories. Partial solutions have indeed already been achieved, as in the unification of electromagnetism and the weak nuclear force.

The pioneers of the first half of this century pointed the way to such developments, and in this respect the most effectual and farsighted may yet prove to have been Eddington.

FURTHER READING

ARTHUR STANLEY EDDINGTON 1882–1944. H. C. Plummer in *Obituary Notices of Fellows of the Royal Society,* Vol. 5, No. 14, pages 112–125; November 1945.

THE DEVELOPMENT AND MEANING OF EDDINGTON'S "FUNDAMENTAL THEORY." Noel B. Slater. Cambridge University Press, 1957.

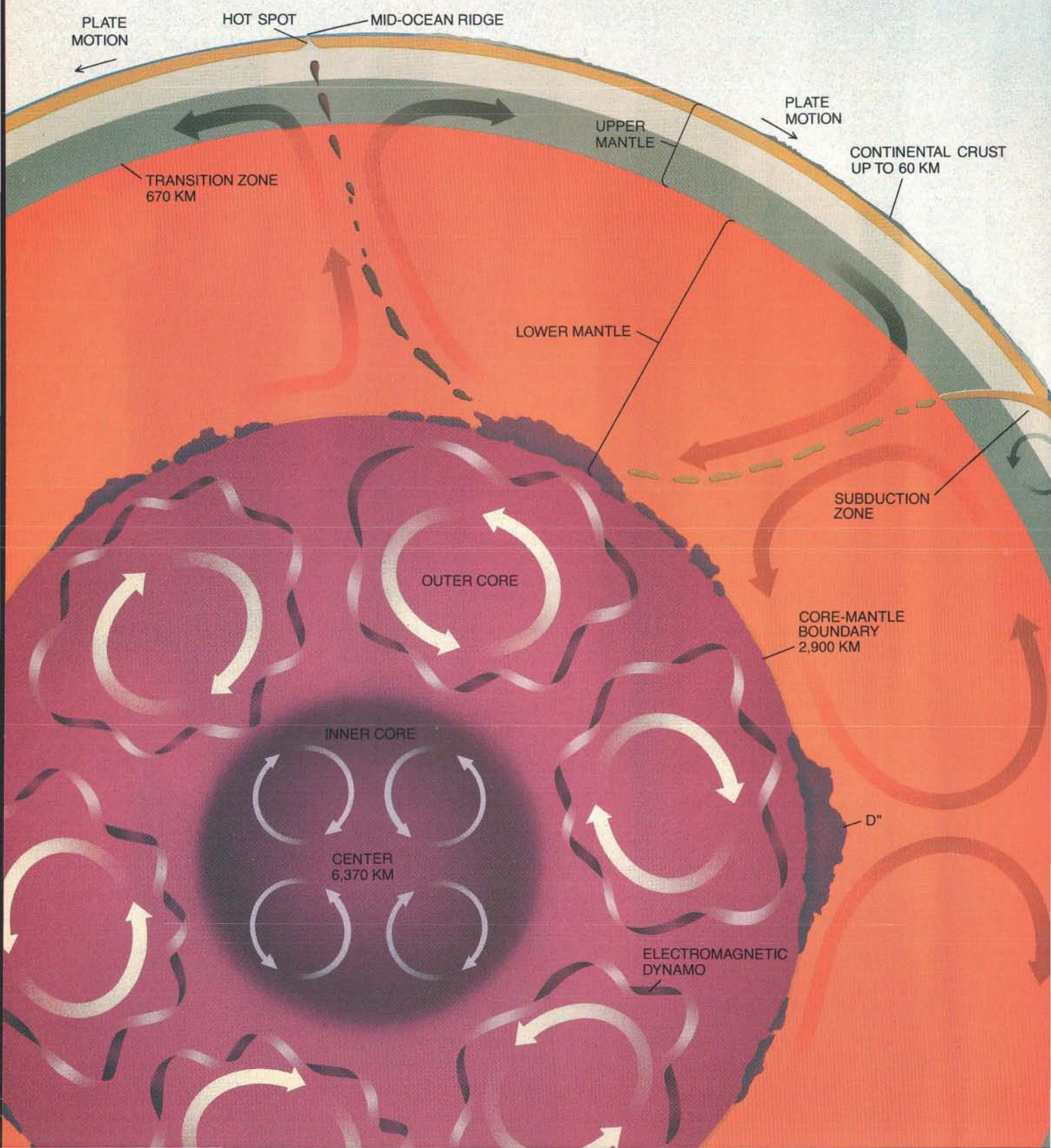
THE LIFE OF ARTHUR STANLEY EDDINGTON. A. Vibert Douglas. Thomas Nelson and Sons, 1957.

MEN OF PHYSICS: SIR ARTHUR EDDINGTON. C. W. Kilmister. Pergamon Press, 1966.

EDDINGTON: THE MOST DISTINGUISHED ASTROPHYSICIST OF HIS TIME. Subrahmanyan Chandrasekhar. Cambridge University Press, 1983.

PEERING INWARD

by Corey S. Powell, *staff writer*

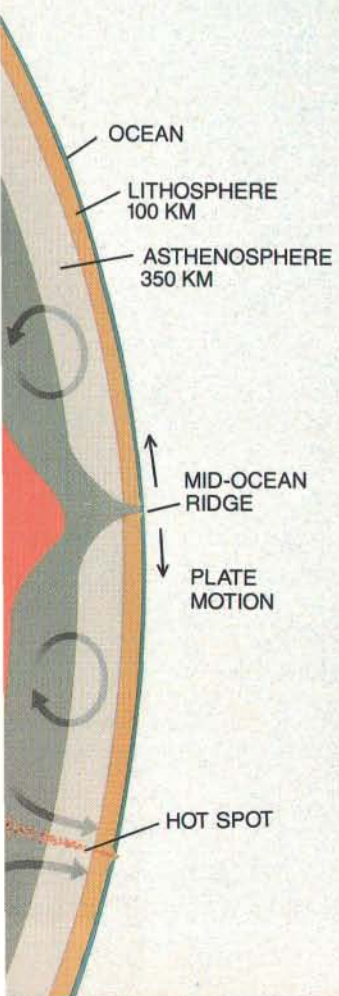


New technologies enable geophysicists to discern the structures hidden deep within the earth. Their findings even offer a glimpse into the future of the planet.

I've looked thousands of kilometers through the inside of the earth," says Gary A. Glatzmaier nonchalantly. While the geophysicist from Los Alamos National Laboratory sits securely on the surface of the earth, an ultrapowerful supercomputer enables him to simulate the conditions deep in the interior. The world that he sees there churns restlessly like a boiling cauldron as it tries to shed its internal heat. Vast sheets of cool rock break free and sink slowly into the hot interior. Plumes of hot, molten rock rise and expand like mushroom clouds as they near the surface.

Glatzmaier is part of the geophysical vanguard whose work reveals the earth's complex internal dynamics. These researchers look beneath the thin outer skin (no more than 60 kilometers thick) that contains all the familiar uplifted mountains, shifting faults, erupting volcanoes and the like. By probing the underlying layers of core and mantle that make up more than 99 percent of the planet, they are uncovering the processes that ultimately shape the surface.

Glatzmaier's simulations show the currents in the rocky but pliable mantle, which extends 2,900 kilometers down. In the real world, these currents move at an extremely leisurely pace—taking hundreds of millions of years to complete one loop—but they are powerful enough to move continents and reshape oceans. Activity also occurs deeper down, in the earth's iron-alloy core. Agitated flows in the outer, liquid part of the core, which move a million times more briskly than those in the mantle, generate the earth's protective magnetic field. Recently geophysicists have begun to chart the circulation of this subterranean metallic ocean. Sophisticated seismic techniques even permit a peek into the solid inner core, whose outer edge lies more than three fourths the way down the earth's 6,370-kilometer radius.



CEASELESS CHURNING of the earth results as heat tries to escape from its interior. Twisted currents in the molten iron-alloy outer core generate the earth's magnetic field. Powerful but leisurely convective motions in the silicate mantle move continents and drive nearly all surface geologic activity. Material may circulate all the way from the lowermost part of the mantle (the D″) to the top of the mantle, in which case plumes of hot rock could drag bits of the core to the surface (top regions). *On the other hand, a sudden change in mineral structure and composition 670 kilometers below the surface may cause layered convection* (right regions). *Numbers indicate depth below the surface of the earth.*

This new picture of the earth's inner structure is all the more remarkable given how little was known until recently. The earth's core—which is more than half the diameter of the entire globe—was discovered only in 1906. Moreover, the initial pace of discovery was painfully slow. Sixty-five years elapsed before geophysicists could say with conviction that the core is divided into an outer, liquid part and an inner, solid part. The theory of continental drift, which intimately intertwines with the concept of large convection motions in the mantle, did not gain a firm footing until the 1960s.

Progress in three general areas accounts for the rapidly sharpening picture: growing computer power; more thorough collection of gravitational, magnetic and seismic data; and the advent of devices that can re-create the temperatures and pressures in the center of the earth. Researchers are reading earthquake waves and analyzing the earth's magnetic field to map both the core and mantle. Using sophisticated diamond anvil chambers, they are studying the behavior of rocks at the hellish temperatures and pressures found at the center of the planet.

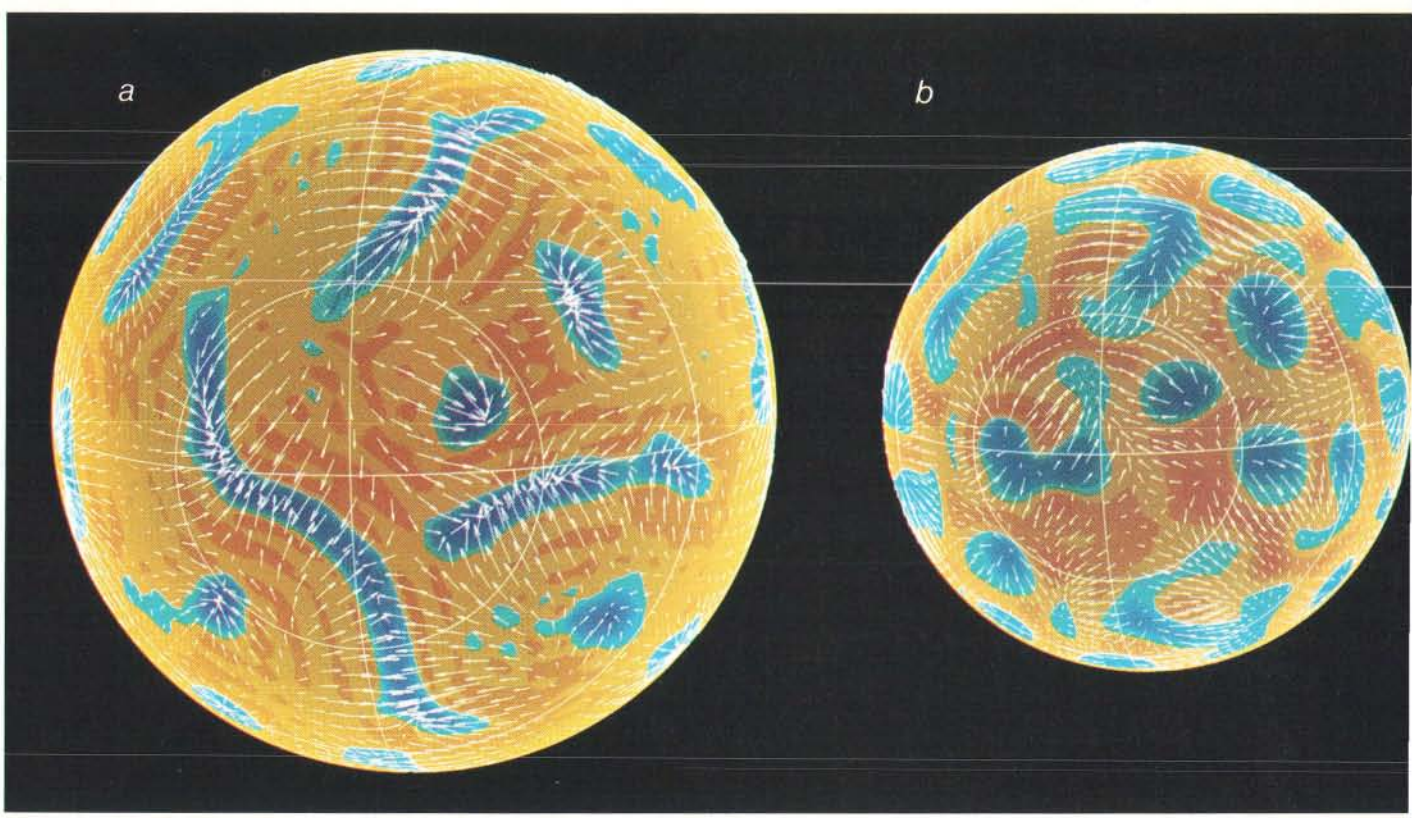
As pieces of information begin to fit together, geophysicists are finding that they not only can reconstruct the behavior of the present earth but also can gain insight into the earth's evolution—and even its ultimate fate, billions of years into the future. "It's a very exciting time," says Adam M. Dziewonski. The walls of his office in the Hoffman Laboratory at Harvard University are cluttered with maps of the earth's interior and the locations of earthquakes, but he keeps a space free for a projector. He runs through a series of slides to make his point.

The pictures on the screen show fields of amorphous red and blue shapes that broadly resemble the model earths created by Glatzmaier's team. In this case, however, the images are not computer simulations but reconstructions of the interior of the real earth. They are the product of an innovative technique called seismic tomography, which several researchers pioneered early in the 1980s [see "Seismic Tomography," by Don L. Anderson and Adam M. Dziewonski; SCIENTIFIC AMERICAN, October 1984].

Seismic tomography takes advantage of the fact that earthquake waves move at different velocities in different parts of the earth. Density, composition, mineral structure and degree of melting all affect wave speed. Temperature is another important factor: all else being equal, seismic waves travel fast in cold regions, more slowly in hot ones. Detecting and analyzing seismic waves from around the globe make it possible to trace a large number of paths through the earth.

Computer analysis of this information enables researchers to convert seismic waves from earthquakes into three-dimensional images of the interior. Dziewonski's maps use the color scheme that has become de rigueur for geophysicists: red denotes relatively slow seismic waves, corresponding to high temperatures; blue regions indicate fast waves, or cool temperatures.

To depths of 200 kilometers or so, the earth appears cold under continents and hot below mid-ocean ridges, where lithospheric plates are moving apart. Farther down, the structure of the mantle seems to correlate not with specific regions of geologic activity but rather with the large-scale motions of the continents. Hot material underlies Africa and the middle of the Pacific Ocean. A ring of cold rock surrounding the Pacific sits under all the continents except Africa, as if they have



HOT AND COLD FLOWS rise and sink through the mantle. The first three images are snapshots of a three-dimensional model of the convecting earth. Blues represent colder than average material; orange and yellow, hotter than average. Arrows indicate the surface flow, at speeds of no more than two centimeters per year. The globes show spherical slices through the model earth at depths of 430 kilometers (*a*), 2,020 kilometers (*b*) and 2,600 kilometers, just above the

been pulled to their present locations by sinking flows in the mantle.

The results of tomography seem to corroborate the intuitive notion that mantle convection determines the motion of the tectonic plates. Despite such conceptual advances, however, the study of deep-earth geophysics remains animated by intense, if generally polite, debate. Nowhere is the discussion more lively than among geophysicists pondering the intricacies of the boundary between the core and the mantle.

Studying the Segregated Earth

The core-mantle boundary or, to save one's breath, the CMB, represents the most abrupt physical and chemical transition inside the earth. Here, silicate rock in the mantle confronts the iron-alloy core. Seismic readings indicate that the rock at the bottom of the mantle is solid and flows extremely slowly, somewhat like glass. In contrast, the outer part of the core has roughly the consistency of water. The temperature difference between the mantle and the core may be as much as 1,000 degrees Celsius.

The CMB probably harbors valuable clues about how the earth functions and how it has evolved. But finding out what goes on at the CMB has bedeviled researchers. Tomography reveals only the way that the earth's inner layers affect the motion of seismic waves, and the resolution of most seismic techniques is poor at the CMB. "There have been a lot of wild ideas floating around because we couldn't test them," muses Don L. Anderson, a professor of geophysics at the California Institute of Technology and one of the pioneers of seismic tomography, adding, "—until now."

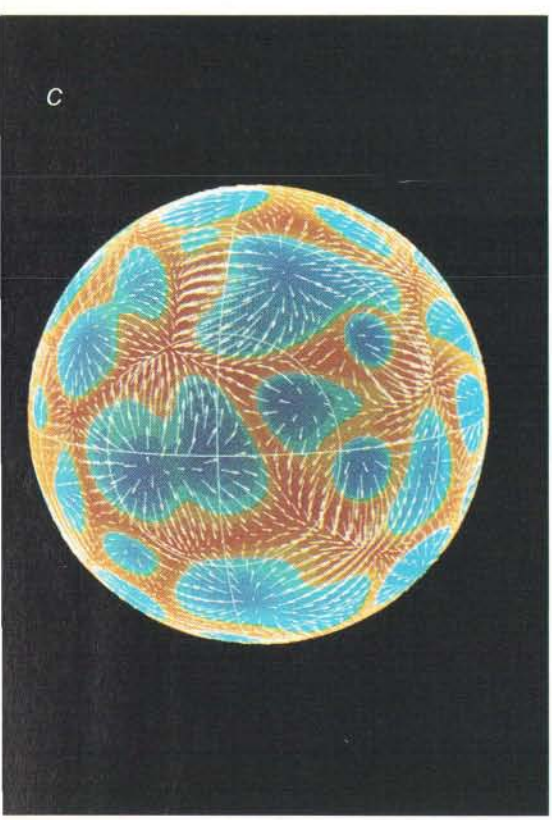
Although scientists cannot go to the core, they can now, in effect, bring the core into the laboratory by means of high-pressure research tools, in particular the diamond anvil cell [*see box on page 79*]. Researchers at the Carnegie Institution's Geophysical Laboratory in Washington, D.C., the University of California at Berkeley and a handful of other locations can create for extended periods the physical conditions at the center of the earth, albeit on a very small scale (samples are no more than one tenth of a millimeter across).

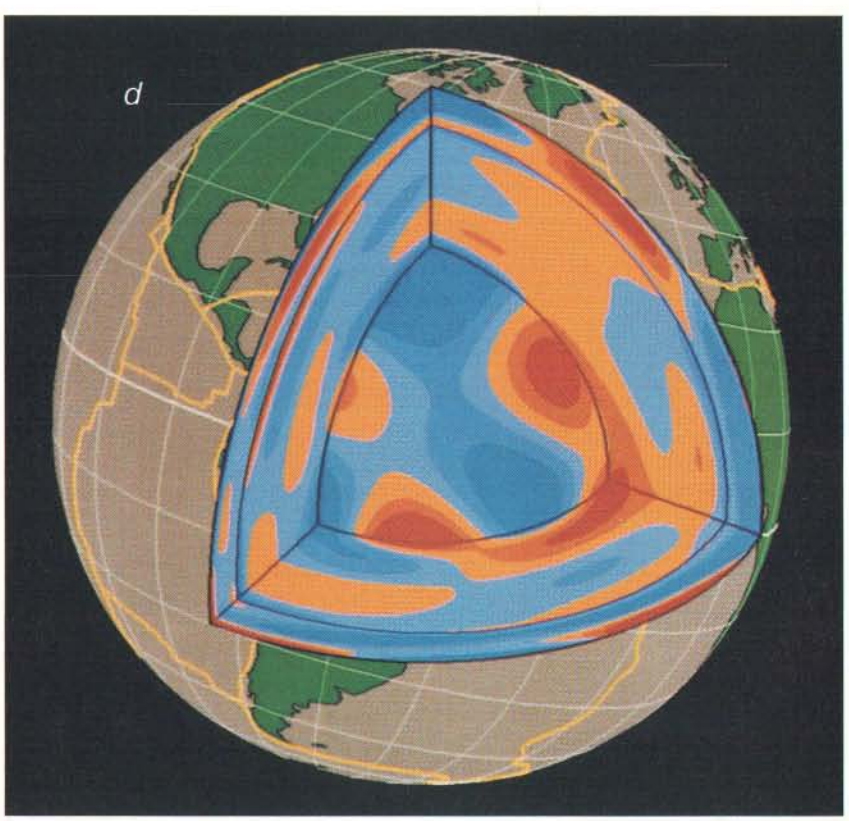
These high-pressure studies are crucial to deciphering the complex chemistry that should occur at the CMB. With temperatures hovering anywhere from 3,000 to 4,500 degrees C and the overlying earth bearing down with more than one million times the pressure at sea level, compounds interact in unfamiliar, and still poorly understood, ways.

One of the most active researchers trying to use experimental evidence to learn about the CMB is Raymond Jeanloz of Berkeley. Jeanloz sports a 1960s-style ponytail and a self-professed enthusiasm for "groping around" at the edge of established geophysics. Admirers and skeptics alike sometimes consider his notions a little "far out," but nobody denies his role in stimulating provocative discussions. David J. Stevenson of Caltech succinctly sums up Jeanloz as a "reconnaissance experimenter."

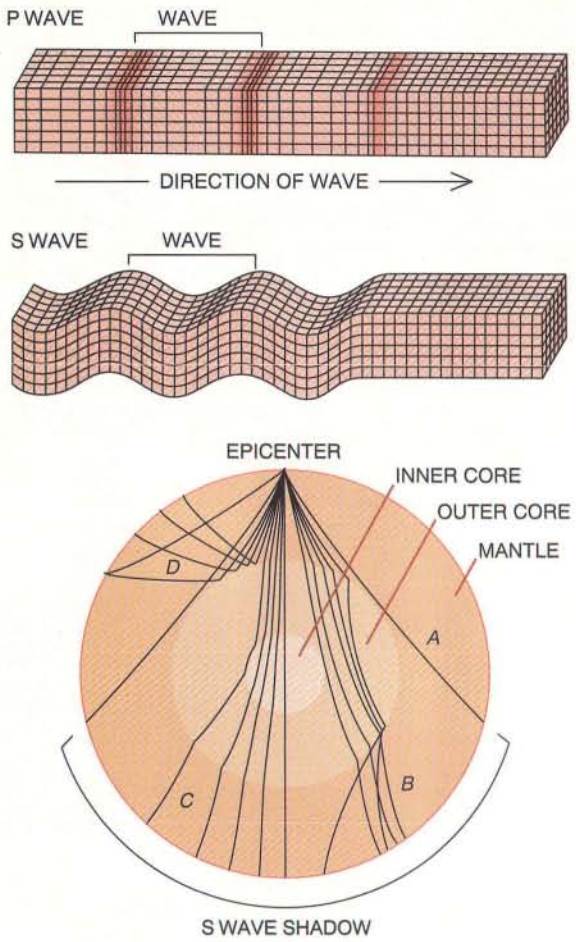
Stevenson refers specifically to Jeanloz's current work on deducing the chemical reactions that occur at the CMB. Jeanloz and Elise Knittle, who is now at the University of California at Santa Cruz, have suggested that the CMB is a leaky barrier that permits considerable chemical interaction between the mantle and the core.

Jeanloz and Knittle base their conclusion on an experiment in which they





core-mantle boundary (*c*). Cold, downward flows dominate the convective process; hot, displaced material spreads out as it rises. Tomography yields a three-dimensional map (*d*) of seismic velocities from the top of the mantle down to the core-mantle boundary. Slow (*red*) and fast (*blue*) regions roughly correspond to hot, rising and cold, sinking material, respectively. Flow patterns in the upper mantle appear appreciably different than those farther down.



How Seismic Waves Reveal the Earth's Inner Workings

Seismic waves set off by an earthquake or large explosion serve as probes of the earth's internal structure. P waves consist of compressional pulses through the earth, analogous to sound waves in air. They can propagate through every part of the earth's interior. S waves are transverse deformations of the solid earth. They can travel only through a resilient material and so cannot pass directly through the liquid outer core. The British seismologist R. D. Oldham inferred the existence of the core in 1906 by noticing its seismic S wave shadow.

Variations in temperature, pressure and composition within the earth alter the wave speed and cause seismic waves to bend or even to reflect. Reflections are particularly strong at the surface, the core-mantle boundary (CMB) and the top of the solid inner core. Studying the travel times of waves taking different paths through the earth therefore enables seismologists to deduce the three-dimensional physical properties of the deep earth.

Some waves pass straight through the mantle (*A*), the mantle and outer core (*B*) or the mantle and both the outer and inner core (*C*). Waves that reflect one or more times inside the earth are proving especially useful for mapping the interior. In one technique for examining structures at the bottom of the mantle, travel times of S waves that reflect off the CMB (*D*) are subtracted from those of S waves arriving at the same location via a direct path. Comparisons of P waves reflected off the top of the core with those passing through imply that the height of the CMB varies by as much as a few kilometers from place to place.

loaded a diamond anvil cavity with bits of silicate minerals (the best guess for the composition of the lower mantle) and iron (to simulate the outer core). They then heated them to CMB-like temperatures, accompanied by pressures of up to 800,000 atmospheres.

Afterward, they examined the tiny sample and found evidence that silicon and iron, which remain chemically aloof at the surface, intermix in that environment. Thus, silicon or oxygen from the mantle could find its way into the outer core by forming alloys with iron. These alloys would gradually change the composition and reduce the density of the overall core.

In fact, because the density of the core is noticeably less than that of a pure nickel-iron mixture, geophysicists generally agree that some lighter component must also be present. Most researchers have assumed that the impurities have been there pretty much since the formation of the earth. Jeanloz's added twist is that the composition of the core may be evolving continuously as bits of silicate mantle dissolve into it.

Under the right conditions, material from the mantle that has mixed into the core could form a distinct layer on the core's surface. Thorne Lay, a seismologist at Santa Cruz, perceives a drop in wave velocities at the top of the core. He suggests he may be seeing an effect of the silicon-rich flotsam on the core.

The idea that the core may consist of distinct layers might solve another geophysical puzzle. Studies of the behavior of the geomagnetic field indicate a smooth flow at the surface of the core, whereas seismic studies by Dziewonski and others suggest the presence of several kilometers of topography at the CMB. Perhaps a bumpy layer of core-mantle mixture rides atop the core's otherwise even surface.

Then again, layering of the core may yet prove to be a phantom. One floor above Dziewonski's office at Harvard, Jeremy Bloxham voices his skepticism of this "ad hoc" explanation. After years of correlating historical magnetic field data, Bloxham reports that "it is becoming ever more difficult" to reconcile his findings with the existence of a differentiated top layer.

On the other side of the CMB, the plot grows thicker. Mounting evidence points to the presence of mysterious structures just above the CMB, a region known as the D″ (pronounced "D double prime"). A number of studies, including ones conducted by Lay, reveal that the speed of seismic waves changes some 200 to 300 kilometers above the CMB, indicating the presence of a transitional layer. Other seismic investigations, however, show a nearly clean CMB. "The party line," Lay says, "is that there is a strongly varying structure at the CMB," present in some locations but not in others.

But what are these structures? Suddenly the party line disintegrates. "There's almost no end to the theories," Lay notes. Jeanloz proposes that what goes down must come up: just as mantle may dissolve into the core, material from the core may escape upward into the mantle. He throws out the idea that capillary action could draw iron upward from the core into the mantle, where it would form aggregations of iron alloys or dense iron-enriched silicate rock. Local variations in temperature or chem-

ical composition would explain the blotchiness of the patches at the D″.

While not disputing that the D″ may receive material from the core, Stevenson points out other possible sources of intermixing. He proposes that the bottom of the mantle "could also be a junk pile of stuff from the mantle." This layer could be the dregs of old subducted lithospheric plates, for example. The D″ may also mark the resting place for primordial material that sank through the mantle but was too light to join the core. Many of his colleagues echo this notion.

Lay adds that temperature and pressure effects probably significantly affect the physical structure of the D″. Because of the extreme temperature gap at the CMB, heat flowing from the core could cause the D″ to become periodically unstable. Hot, buoyant material from the D″ might then form a hot, rising plume. Peter Olson of Johns Hopkins University and a handful of others have developed models showing that mantle plumes should take the form of narrow, focused conduits of hot material. Where such a conduit reaches the surface, it might create a volcanic hot spot.

Hot spots are volcanically active regions that appear to remain fixed for many millions of years as the lithospheric plates bearing the continents slide by. A hot spot beneath the Pacific Ocean created the neat line of Hawaiian Islands; its presence can still be seen in the frequent eruptions of Mount Kilauea. Evidently, the spots originate from a stationary, and hence extremely deep, source that burns through the seafloor or continents as they slowly drift across the mantle.

One Mantle or Two?

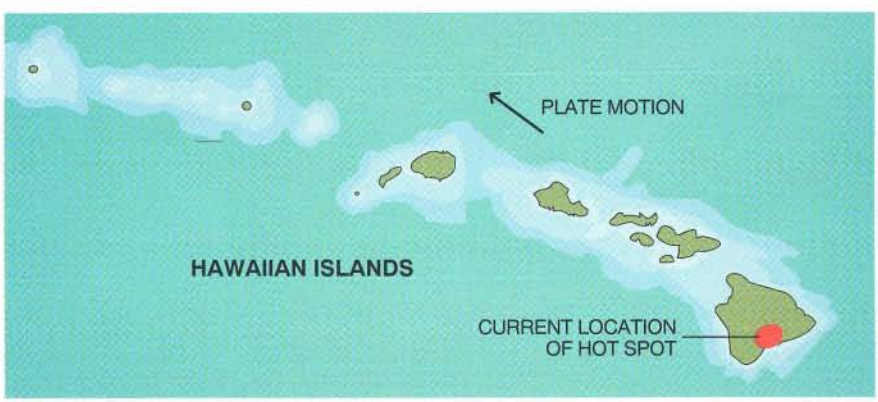
If plumes originate at the CMB, they might carry traces of core material all the way to the surface. Gases expelled by surface volcanoes contain helium 3, a relic from the formation of the earth. Some of this gas could have been trapped for billions of years in the mantle, but Stevenson speculates that some may also have leaked from the core into the D″, where it was dragged upward. "There is no way to prove it," he admits, "but it seems reasonable that some atoms at the surface of the earth have been in the core recently in geologic time."

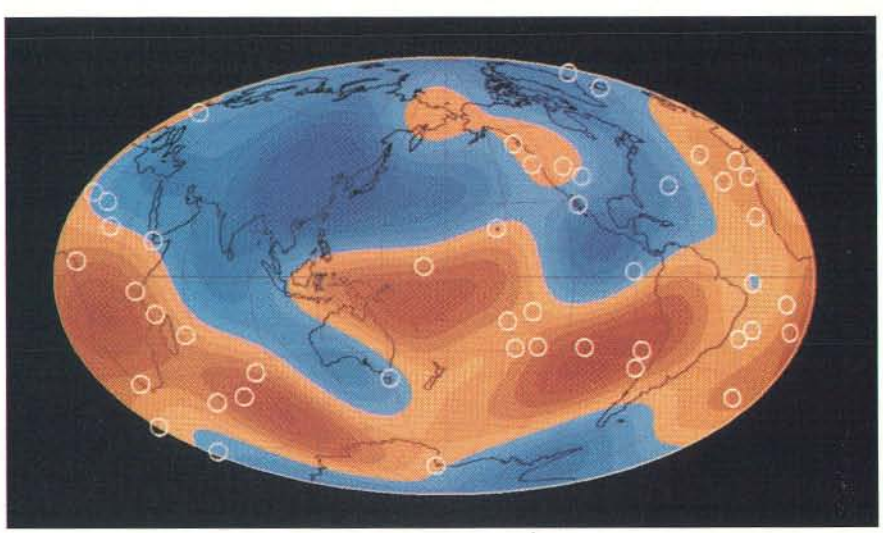
Anderson, on the other hand, has a very different perspective on the fate of the material at the D″. He argues that the material at the bottom of the mantle is too dense to ascend all the way to the crust. Small convection cells might form in the D″. If the layer gets hot enough, it might rise and then sink back through the lower mantle, "but this is a very different story than plumes rising to the surface," he says.

The way that plumes travel through the earth depends, of course, on how the mantle circulates. And this, too, is a subject of considerable debate. One school of thought holds that the entire mantle mixes together via huge convection currents that extend from the CMB all the way to the bottom of the crust. The other camp favors a mantle composed of at least two distinct, independently convecting layers. Overall, Stevenson guesses that "about 80 or 90 percent" of geophysicists fall on the side of deep-mantle convection. But Russell J. Hemley, a high-pressure physicist at the Geophysical Laboratory, thinks the tide is going the other way because of growing experimental evidence that the mantle's composition changes significantly with depth.

Many researchers cite the distribution of hot spots as an argument in support of deep convection. Working with Mark A. Richards of Berkeley and Bradford H. Hager of the Massachusetts Institute of Technology, Dziewonski has shown that hot spots on the surface seem to occur preferentially over regions at the CMB that appear relatively hot in seismic tomographic maps. The provocative implications are that hot spots siphon off these temperature excesses and that a direct link exists between the CMB and surface volcanoes.

Computer simulations developed by Glatzmaier and fellow computer-modeling pioneers David Bercovici of the University of Hawaii and Gerald Schubert of the University of California at Los Angeles suggest that internal heat sources (such as radioactive decay) drive large-scale convection, whereas heating from below is more likely to





MANTLE HOT SPOT burned through the earth's crust, creating the linear chain of Hawaiian Islands (*top*). Hot spots remain stationary while the overlying lithospheric plates slowly slide by. A map of seismic velocities at the base of the mantle (*bottom*) indicates that nearly all hot spots (*superimposed white circles*) sit above relatively high-temperature regions in the lower mantle (*red colors*), hinting that hot spots might originate from a layer just above the earth's hot core.

produce plumes. Based on such simulations, Geoffrey Davies, a hot-spot expert at Australian National University, suggests that mantle convection vents the heat of the mantle and that hot-spot plumes siphon off the heat of the core. If plumes did not come from the CMB, he argues, heat rising from the core would create strong upwelling currents, which are not in fact observed.

Thomas Jordan, head of the earth, atmospheric and planetary sciences department at M.I.T., also sees evidence that convection extends continuously through the mantle. Jordan shuns tomographic techniques and concentrates instead on studying seismic waves that have bounced off the CMB to map the fine structure of the mantle. From these studies, he envisions a "conveyor belt system" in which huge convective cells move the continents at the surface and, in a peculiar kind of parallel world, move around the irregular lumps of material above the CMB.

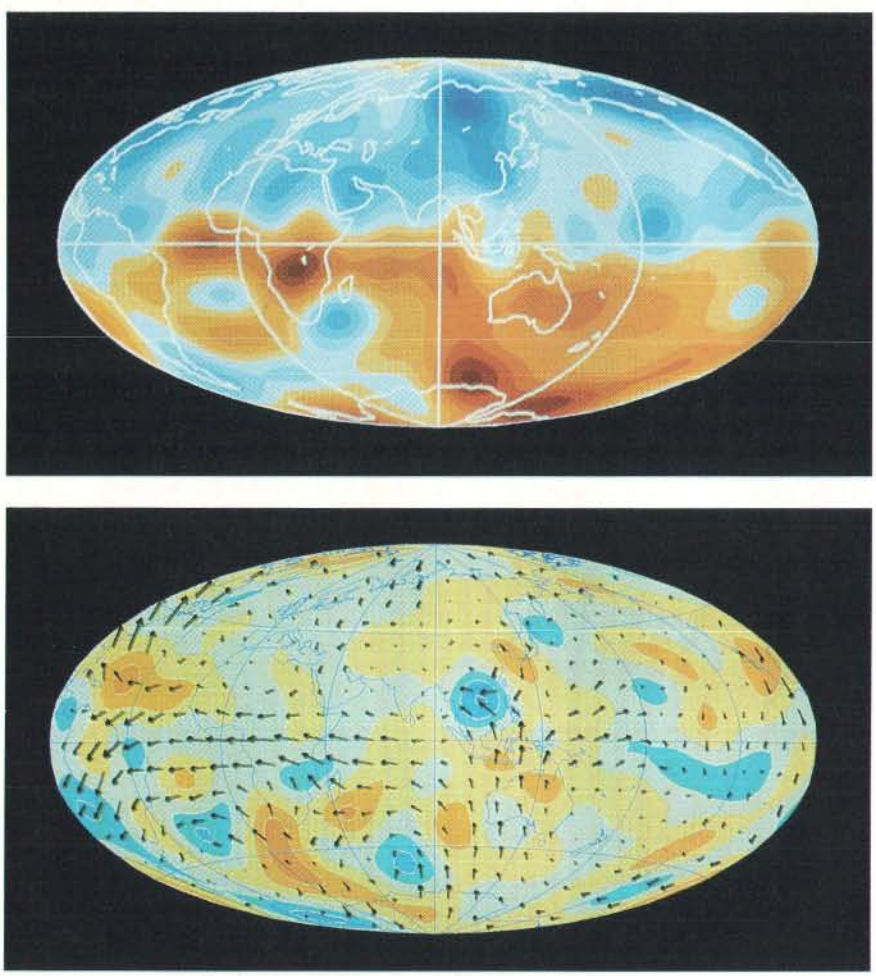
Doubters of the deep-circulation model see things a bit differently. Jeanloz and Knittle's experiments imply that the lower mantle is too dense for it to have the same composition as the upper mantle. Injecting some iron from the core into the lower mantle would increase the density to the observed levels. Jeanloz concludes that deep convection does not occur, because it would disrupt this chemical segregation. Moreover, dense material in the lower mantle would never be able to rise very far.

Other researchers are not persuaded by that line of reasoning. Groups at the State University of New York at Stony Brook and the Geophysical Laboratory find that perovskite, the dominant mineral in the lower mantle, is significantly more dense at high pressures than Jeanloz's experiments indicate. Based on his own diamond-cell investigations of mantle material, Hemley thinks Jeanloz's work on mantle density "needs to be modified."

The behavior of the mantle seems to become especially complex in the so-called transition zone between 400 and 670 kilometers beneath the surface, where the speed of seismic waves changes abruptly. The depth of this zone appears to be determined by the pressures at which minerals in the upper mantle (primarily the mineral olivine) deform to new, quasistable states (perovskite and wüstite). Lithospheric plates sucked down into the mantle often appear to bend and deflect when they reach the transition zone, as if they have hit an impenetrable boundary.

Supporters of layered convection hold that one loop of mantle convection travels between the CMB and the 670-kilometer transition zone, the other between the transition zone and the top of the mantle. Material from below 670 kilometers would not reach the surface, and events at the CMB would only indirectly affect the upper mantle.

Anderson perceives "no evidence" that any material below 670 kilometers is directly involved in surface phenomena. He points out that it should be much easier for plumes to form at 670 kilometers, where they can tap into heat flowing from most of the mantle, rather than at the CMB, where the core is the only source of heat. He also notes that the compositions of lavas from different volcanoes indicate that the mantle is not well mixed, so there is no need to invoke material dredged up from the core.



HOT AND COLD FLOWS rise and sink through the mantle. The first three images are snapshots of a three-dimensional model of the convecting earth. Blues represent colder than average material; orange and yellow, hotter than average. Arrows indicate the surface flow, at speeds of no more than two centimeters per year. The globes show spherical slices through the model earth at depths of 430 kilometers (*a*), 2,020 kilometers (*b*) and 2,600 kilometers, just above the core-mantle boundary

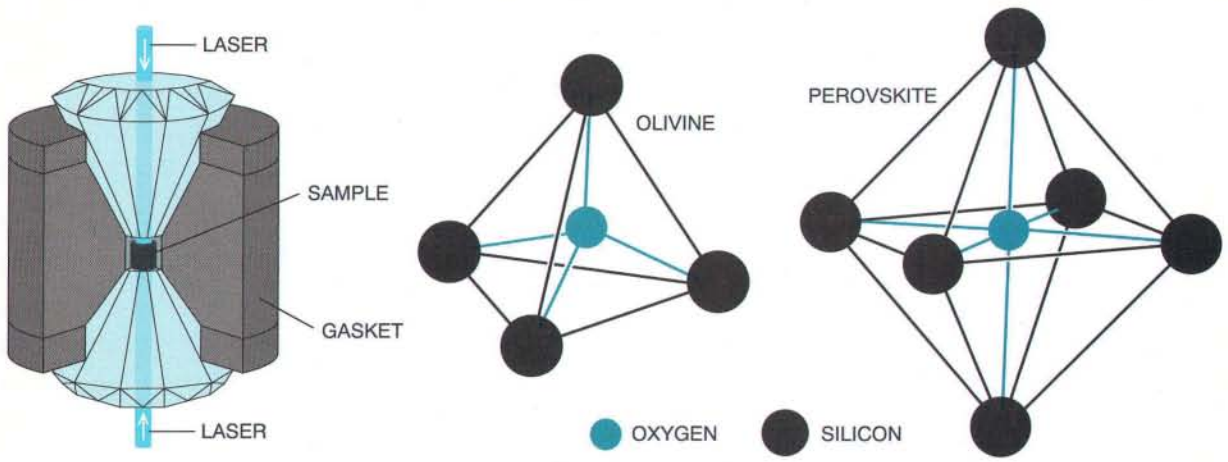
Charting a Middle Path

But deep-earth geophysics remains a very tentative field, and what at first appear to be bitter conflicts often evaporate into mild disagreements of interpretation. Lay points out that hot regions in the lower mantle could act like a "hot plate," producing corresponding hot regions in the upper mantle even if the two do not intermix. Jeanloz earnestly concedes that "the whole-mantle convection models might be right." Most likely, he thinks, "the real world is somewhere in between." For instance, the mantle may be chemically stratified in a way that permits some material to rise above and sink below the boundary 670 kilometers below the surface.

Philippe Machetel and Patrice Weber of Groupe de Recherche de Géodésie Spatiale in Toulouse, France, have developed a detailed computer model to explore this possibility. Unlike the three-dimensional model earth devel-

Geophysicists Learn to Work under Pressure

High-pressure devices make it possible to simulate in a laboratory the conditions in the mantle and even in the core. The diamond anvil chamber (*left*) incorporates two finely cut diamonds whose parallel faces are only about one tenth of a millimeter across. Squeezing the diamonds together exposes a sample to as much as four million times normal atmospheric pressure. Because diamonds are transparent, it is possible to examine the sample while it is under pressure. Diamonds also permit experimenters to heat the specimen using lasers or to probe its structure by bombarding it with an intense X-ray beam. Researchers are trying to reconcile their laboratory experiments with the observed seismic properties of mantle rocks. The upper mantle consists primarily of minerals having an olivine structure, in which four oxygen atoms surround a silicon atom (*center*). High-pressure experiments have revealed that olivine rearranges into a denser perovskite structure (*right*) at depths greater than about 670 kilometers. In fact, the majority of the earth's bulk probably consists of perovskite minerals.



oped by Glatzmaier, Schubert and Bercovici, the French group can look only at a two-dimensional slice. By restricting themselves to two dimensions, however, they freed up enough computer power to explore more of the physical properties of the real mantle, including those that are affected by changes in temperature or pressure. Using plausible equations for the behavior of mantle minerals at the transition zone, Machetel and Weber see material accumulating 670 kilometers below the surface and then, when a critical state is reached, passing up or down across the barrier [*see illustration on next page*].

The manner in which the mantle convects has broad implications for the chemical evolution of the earth. Hager suggests that the deep mantle is extremely viscous, so that any deep mixing must take place very slowly. Using data such as the rate of heat flow through the mantle and the drag on lithospheric plates, Hager and his colleagues calculated that the lower mantle is roughly 30 times as viscous as the upper mantle.

If Hager is correct, the deep mantle may circulate sufficiently slowly—perhaps once every billion years—that large pockets of mantle remain fairly undisturbed and so still retain much the same composition they had when the earth formed 4.5 billion years ago. These pockets could be the "primitive reservoir" that Stevenson invokes to explain the puffs of helium 3 that emerge from surface volcanoes. Indeed, Anderson suggests that localized sources in the mantle could easily account for the amount of observed helium 3. Stevenson worries, however, that mantle convection was more vigorous in the past, when the earth was hotter, the mantle less viscous and the escape of heat more rapid.

The Realm of the Core

The earth's ceaseless efforts to shed its internal heat affects the innermost world—the dense iron-alloy core—just as it affects the world at the surface. On the top side, the circulation of the outer core is affected by hot and cold regions in the relatively stationary lower mantle. The liquid outer core flows several kilometers a year, about a million times faster than the overlying mantle. At its base, the outer core absorbs heat from the enigmatic inner core.

As the outer core slowly cools, iron crystals freeze out of the molten liquid and settle atop the inner core, releasing heat as they do so. Thus, the inner core has been growing continuously throughout the history of the earth. Despite its great temperature, roughly as hot as the surface of the sun, the enormous pressures at the earth's center keep the inner core solid. Even here, however, heat gradually leaks outward. Recent tomographic studies confirm that the inner core, like the mantle, maintains a rhythm of slow convection as it cools.

The rapid flow of the outer core is dramatically evident even far outside the earth. Geophysicists generally agree that fluid flow in the outer core generates an electric current and thereby powers the earth's magnetic field. If so, then one can "listen" to the sloshing of the outer core simply by looking at the orientation of a very sensitive compass. In principle, it should be possible to work in reverse and reconstruct conditions at the core by measuring the structure and behavior of the field at the surface.

That is what Bloxham at Harvard set out to do. He examines a world far more active than the one seen by seismologists. Changes in the outer core take place over decades as opposed to tens of millions of years for those in the mantle. During the past few centuries, the main north-south (dipole) component of the earth's magnetic field has grown markedly weaker, and the

whole field has drifted westward. Over longer stretches of time, the field periodically decays and then reasserts itself, north and south poles swapped [see "The Evolution of the Earth's Magnetic Field," by Jeremy Bloxham and David Gubbins; SCIENTIFIC AMERICAN, December 1989].

Bloxham makes a somewhat humble analogy between his work and weather forecasting. Examining the weather at one instant reveals only immediate facts about the temperature, humidity or wind. But watching how weather changes in the course of a few days reveals a wealth of information about the behavior of the atmosphere, the flow of the winds, the movement of storms and so on. Likewise, monitoring gradual changes in the earth's magnetic field makes it possible to infer the climate in the outer core.

To tackle this problem, Bloxham teamed up with fellow magnetic field researchers David Gubbins of the University of Leeds and Andrew Jackson, a former Harvard colleague now at the University of Oxford. For their first step, they pored through records of magnetic field measurements dating back to the late 17th century. They then utilized various mathematical techniques to project these measurements onto the surface of the core. In this way, they produced maps of the magnetic flux at the core-mantle boundary that span nearly three centuries. A second round of computations enabled them to reconstruct the flow at the core's surface that produces the observed magnetic field [*see illustration on page 78*].

Old textbooks often depicted the earth's magnetic field as a huge bar magnet placed slightly off-center inside the earth. Bloxham's maps obliterate that image, replacing it with colorful plots of contour lines, done up in the geophysicists' usual reds and blues. The colors define regions of positive and negative polarity. Even the division between north and south pole is no longer simple.

Bloxham finds that patches of reverse polarity invade both hemispheres. The asymmetric placement of the overall field seems to result from the uneven distribution of these patches of abnormal polarity, or "core spots." These core spots evolve noticeably over just a few decades, reflecting the rapid pace of circulation in the outer core.

But "the eruption of core spots is almost certainly influenced by the location of hot spots in the mantle," Bloxham explains. The mantle moves so slowly that it is essentially motionless from the core's perspective, but hot regions in the mantle could encourage particularly vigorous convection in the underlying core material.

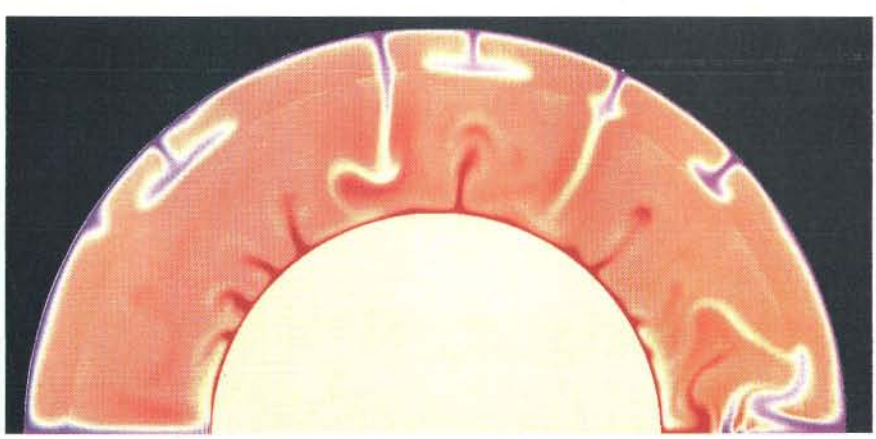
Sure enough, seismic tomographic maps do show apparent regions of hot and cold mantle material immediately above the core, possibly associated with upwelling and downwelling material. Bloxham and Jackson find that some regions of rapid core flow and abnormal polarity—most notably under Africa—seem to correlate with hot regions identified by seismic tomography. Unfortunately, seismic maps become quite uncertain at such great depths, and so any temperature inferences must be taken with a grain of salt.

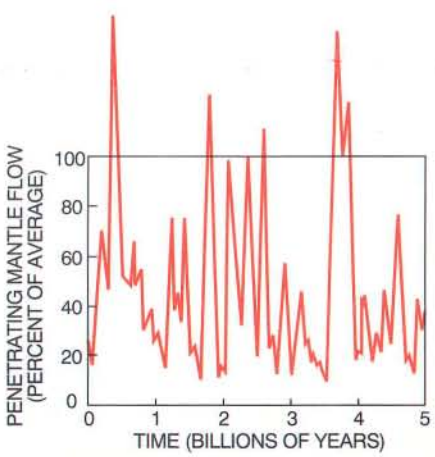
Bloxham realized that by reversing the problem, he could produce his own independent measurements of the temperatures at the top of the core by examining the structure of the field. For example, a spot of relatively hot mantle just above the core should excite upwelling and particularly strong magnetic activity, whereas a mantle cold patch would induce sinking flows and reduced activity. Drawing on this idea, Bloxham and Jackson deduced the temperatures at the CMB. The map that resulted roughly agrees with temperature maps produced by seismic tomography, although many discrepancies show that both approaches still have a long way to go.

Nevertheless, there is more here than smoke and mirrors. Jackson presented a graphic demonstration of the power of magnetic field mapping last winter at the American Geophysical Union meeting in San Francisco. If the magnetic field is produced by core flows, these flows must rub against the inner boundary of the mantle, subtly affecting the rotation of the earth. After whizzing through a number of calculations, he produced a graph of predicted changes in the length of the day. Next he showed a transparency depicting the real, astronomically determined changes, adjusted to remove atmospheric effects. The match was persuasively close.

Reaching below the outer core to examine the inner core involves reverting to the kinds of techniques used for studying the solid mantle. The first glimpses of the inner core have come from tomographic studies that focus on seismic waves that involve pulsation of the entire earth. Earthquakes or large explosions cause the earth to ring like a bell. If the earth were uniform in all directions, every part would oscillate in unison. In fact, the earth is slightly elliptical, and the mantle varies in temperature and composition from place to place, all of which complicate the pattern of oscillation.

After accounting for these factors, there remained some unexplained dis-





INTERMITTENT MIXING of the mantle may explain the competing evidence in favor of both deep and layered convection of the mantle. In this computer model of a two-dimensional slice through the earth, cold, sinking material (*purple*) accumulates at the mineral transition 670 kilometers below the surface (*left*). Occasionally, a nudge from rising return flows lets the cold material break through the transition zone and sink through the lower mantle. Over very long stretches of time, the percentage of the mantle that passes through the transition zone oscillates chaotically (*right*).

tortion in the ringing of the earth. From the detected pattern of oscillations, seismologist John H. Woodhouse of Oxford concluded that the distortion originates in the inner core (a group of seismologists at Caltech disputes that finding, however). He eagerly points to a "pretty obvious explanation" for distortion in the core. Because iron crystals are themselves asymmetric, seismic waves travel faster with the crystals than against them. A preferential alignment of iron crystals in the inner core might explain the observed signal.

But what could cause the crystals to line up? Once again, Jeanloz appears on the scene. While conceding that it is "beyond anyone's ability to make a detailed model of the core," Jeanloz cannot resist speculating that this is an effect of the convection of the inner core. He expects that the motions must be extremely slow, perhaps only a few centimeters per year. Yet such currents could suffice to determine the most likely orientation of iron crystals in the inner core.

Woodhouse agrees that such a convection process seems reasonable. Convection would subtly alter the shape of the inner core (and hence the currents and the crystals), causing it to align itself with the rotation axis. Sure enough, the orientation of the inner core inferred from seismic oscillations lies within 20 degrees of the earth's axis of rotation.

Past and Future Earths

From the inner core all the way to the surface, the dynamic activity of the earth is driven by heat. And, like any heat engine, the earth must be gradually running down.

Predicting the future is a notoriously tricky business, however. In particular, the distinction between deep-mantle convection and layered convection has profound implications for the earth's ultimate fate. An upper mantle that does not mix with the lower mantle could act as an insulating blanket, holding in the earth's heat. Deep convection would bring hot material directly to the base of the crust, permitting the mantle to cool much more quickly. If that is the case, the earth may have changed considerably in its 4.5-billion-year history.

Relics of the earth's past behavior undoubtedly offer clues to its future—if only they can be understood. Norman Sleep, a geophysicist at Stanford University who votes on the deep-convection side, envisions an ancient earth in which the mantle was considerably hotter, less viscous and more vigorous in its circulation. If so, the temperature gap at the CMB would have been far less and hot-spot plumes should have been weak or nonexistent. Sleep observes that, in fact, there is no clear evidence of hot-spot volcanism dating back more than 1.3 billion years.

Anderson looks at the same evidence but reaches a very different conclusion. He claims that it is very difficult to identify hot spots and flood basalts more than a billion years old. Within that age limitation, he sees no sign of a decline in hot-spot activity, consistent with the layered-convection model.

As for the future, Jeanloz, with his usual flair for the dramatic, opines that "the earth will still be active when it is consumed by the sun" in about five billion years. Anderson agrees, observing that the earth has cooled only about 200 degrees C in the past billion years, which he attributes in part to layered convection.

Sleep takes a more pessimistic view. As the mantle cools, it will grow ever stiffer and less mobile. "In the next billion years, the motions of the plates will slow or stop," he predicts with an air of resignation. Continents will settle into permanent positions, and erosion will gradually level the great mountain chains. Hot-spot plumes will continue to wriggle to the surface and will become the primary means by which the earth sheds its internal heat.

Scientists may not have to wait a billion years to test these predictions. They may be looking at the earth of the future right now—on Venus. Images of Venus from the *Magellan* probe reveal a planet dotted with volcanoes but devoid of clear signs of earthlike plate tectonics. Venus, like the future earth, "seems to be dominated by large, sluggish plumes," Sleep says. Ironically, Venus may have cooled faster than the earth because its surface is so hot. Sleep speculates that the high surface temperature on Venus (about 450 degrees C) kept the rocks near the surface soft and pliable, which permitted more thorough circulation of the outer parts of the planet. In this way, Venus was able to rid itself of its internal heat faster than was the earth.

In spite of a remarkable series of technological and conceptual advances that make such speculations possible, deep-earth geophysics is a young discipline and, like any youth, is still plagued with uncertainty. Seismic tomography, which has yielded mind-broadening images of the structure of the earth, has a number of critics. Jordan, for example, voices his opinion that the technique "has run out of steam." High-pressure physicists from Berkeley, Stony Brook and the Geophysical Laboratory question some aspects of one another's work. Researchers debate the merits of two-dimensional computer models versus ones examining all three dimensions. Nearly everyone agrees on the need for better information about the behavior of minerals at high temperatures and pressures.

Thomas Ahrens, who studies the physics of the deep earth at Caltech, takes on the role of general skeptic. He points out that in the lower mantle, various tomographic maps "just don't fit together." Next he tut-tuts about the "wild extrapolations" involved in applying Jeanloz's diamond anvil experiments to the real world. "Models are fine, because they motivate more research," Ahrens comments, noting that the research is desperately needed to resolve the "marked disagreement between the earth and the laboratory."

Nevertheless, the overall tone is one of wonder. "I find it remarkable that we can see into the inner core," Bloxham says. Jordan considers recent advances in the analysis of seismic waves "absolutely revolutionary." Jeanloz uses similar words to describe new high-pressure technologies but sounds equally dazzled by "quantum leaps on the seismology and geomagnetism side."

A series of links now connects drifting continents, earthquakes, the geomagnetic field and volcanoes. Each phenomenon contributes a little to the increasingly acute perceptions of our planet, and each dramatically underscores that the latest geophysical results are meaningful and real. "It's like the old story about the elephant—everyone's looking at a different part of the story," Anderson quips. "But now all the blindfolded Indians are starting to talk to one another."

FURTHER READING

INSIDE THE EARTH: EVIDENCE FROM EARTHQUAKES. Bruce A. Bolt. W. H. Freeman, 1982.

THE DYNAMIC EARTH. *Scientific American* (special issue), Vol. 249, No. 3; September 1983.

GLOBAL IMAGES OF THE EARTH'S INTERIOR. Adam M. Dziewonski and John H. Woodhouse in *Science*, Vol. 236, pages 37–48; April 3, 1987.

STUDIES OF THE EARTH'S DEEP INTERIOR: GOALS AND TRENDS. Thorne Lay, Thomas J. Ahrens, Peter Olson, Joseph Smyth and David Loper in *Physics Today*, Vol. 63, No. 10, pages 44–52; October 1990.

EARTH'S CORE-MANTLE BOUNDARY: RESULTS OF EXPERIMENTS AT HIGH PRESSURES AND TEMPERATURES. Elise Knittle and Raymond Jeanloz in *Science*, Vol. 251, pages 1438–1443; March 22, 1991.

Flat Horizons

U.S. pursues research but little development of advanced screens

Researchers at the IBM Corporation have been developing flat-panel display technologies for more than two decades. In mid-May some of these efforts will at last see light as the company inaugurates its first plant dedicated to manufacturing flat-panel computer screens—the only such facility funded by a U.S. company. IBM owns only 50 percent of the venture, however. The other half is held by the Japanese company Toshiba. Moreover, the plant itself is in western Japan, in the town of Himeji.

The overseas migration of IBM's efforts in flat-panel displays brings into focus a tough issue for the U.S.: manufacturing. "The U.S. doesn't have an R&D problem," observes James C. McGroddy, a vice president and director of research at IBM. "The real issue is manufacturing and the ability to make an investment."

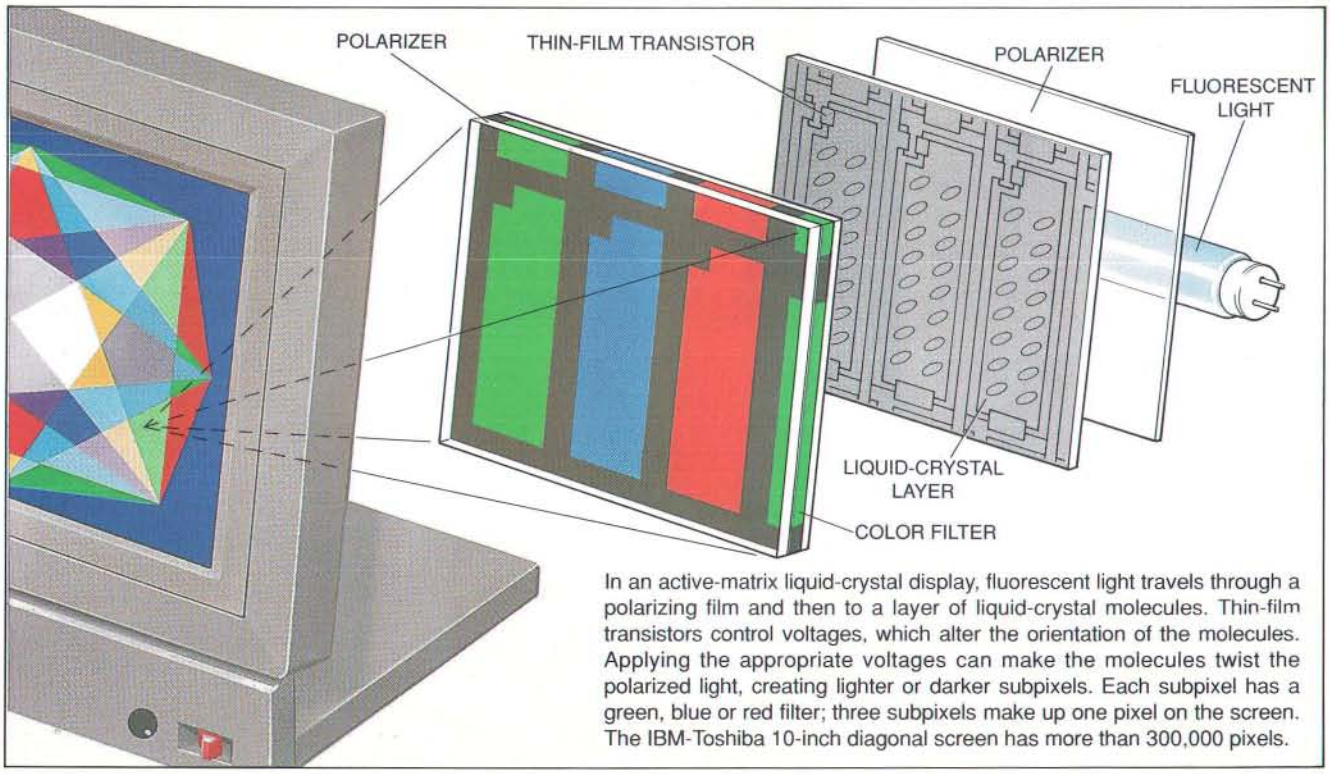
Few experts deny the growing importance of advanced display technologies; their uses range from lightweight flat-panel screens for portable computers to equipment for projecting wall-size high-definition television images. Sales of flat-panel screens primarily for computers and small televisions reached $2 billion in 1990 and will climb beyond $5 billion in 1995, according to Stanford Resources, a market research firm in San Jose, Calif. (The Department of Commerce, moreover, has been reviewing charges that Japanese producers are undercutting nascent U.S. efforts by selling displays at below fair market prices.)

Nevertheless, when IBM began drawing up plans for a plant, McGroddy argued for a location in Japan. He cites three reasons: "There's a stronger infrastructure in Japan. The tool industry, the people who supply the things that we need, such as color filters, these are in Japan. Also the cost of capital is lower." And most of the development of the displays, both by IBM and Toshiba, was done in Japan.

For technologists, designing and building flat-panel and projection displays that have a better picture and a lower price than do bulky cathode-ray tubes are stiff challenges. Companies must develop expertise in many activities, such as materials processing and lithography. To do so requires time and commitment. "Some people in the U.S. seem to think that somewhere, someone will invent a new technology that will leapfrog all the rest of this development," says Lawrence E. Tannas, Jr., a display-technology consultant in Orange, Calif. Instead, he suggests, "the surprises we get are in the continual progress in the existing technologies."

The IBM-Toshiba venture, for instance, builds on work pioneered in the 1970s at RCA and Westinghouse on a flat-screen approach known as active-matrix liquid-crystal displays. The key to such displays are organic liquid-crystal molecules, which can be realigned by applying an electric field. Depending on their orientation, these molecules partially block incoming polarized light. The light then passes through a color filter and emerges as a single shade of red, blue or green. Combining three colored dots produces one colored picture element, or pixel, on the display.

The orientation of the liquid-crystal molecules is controlled by a matrix of thin-film transistors etched on a plate of glass or amorphous silicon. Every



In an active-matrix liquid-crystal display, fluorescent light travels through a polarizing film and then to a layer of liquid-crystal molecules. Thin-film transistors control voltages, which alter the orientation of the molecules. Applying the appropriate voltages can make the molecules twist the polarized light, creating lighter or darker subpixels. Each subpixel has a green, blue or red filter; three subpixels make up one pixel on the screen. The IBM-Toshiba 10-inch diagonal screen has more than 300,000 pixels.

200-square-micron area of molecules (contributing one third of a pixel) is controlled by one transistor.

Making these transistor arrays, however, is painstaking work. The fabrication techniques are much the same as those used to pattern integrated circuits. But manufacturers have had little experience working with the substrates needed for larger display screens such as 14-inch diagonal plates. Some experts estimate that only 20 percent of every batch of manufactured active-matrix liquid-crystal displays are free of defects.

Yet because active-matrix liquid-crystal displays are the leading candidate for many small screens, Japanese firms are pouring money into building new factories. "It's like an adolescent growth spurt," Tannas observes. Intimidated by the heft of Japan's investment, U.S. firms are hoping to build larger screens and exploit different approaches.

One unexpected developer of such displays is Xerox Corporation. Researchers at the company's Palo Alto Research Center had been etching thin-film transistors on both amorphous silicon and polysilicon for use in their advanced printers and scanners, recalls Malcolm J. Thompson, who manages the electronics and imaging laboratory. When other Xerox researchers requested a large panel display that could be used as an interactive computer window in a meeting room, Thompson's team added the necessary layer of liquid-crystal molecules on top of the thin-film transistors and within 18 months produced a prototype.

Unlike Toshiba and IBM, Xerox aims to build large direct-view and projection displays primarily from polysilicon. Although more difficult to make than amorphous silicon films, polysilicon-based displays promise more functionality.

By mid-May, Thompson hopes to have in working order a 13-inch polysilicon system. Xerox is eager to find a strong manufacturing partner that will turn the prototypes into products. Even though most potential partners are not U.S.-based firms, Thompson reports a "quite high probability" that a joint manufacturing operation would be located in the U.S.

One way to avoid the problems of etching transistors on large plates may be to build a mosaic of small active-matrix liquid-crystal modules. T. Peter Brody, a pioneer of these displays and a co-founder of Magnascreen Corporation in Pittsburgh, realized that if the pixels on the edges of adjacent modules are no farther apart than the pixels on the modules themselves, he could create a large, seamless display panel. But to tile the modules so closely together, the researchers had to turn to a recently developed material: polymer-dispersed liquid-crystal films.

Progress is slow; Magnascreen has successfully assembled only four modules, each measuring three by four inches. Moreover, disagreements over how the company should try to commercialize the work forced Brody to leave the firm, he says. "I felt the technology had to be done on a big scale, not for a niche market."

Others are trying to alter the intensity of light passing through a panel with gas or with mirrors. Instead of using transistors to twist the liquid-crystal molecules, workers at Tektronix Laboratories in Beaverton, Ore., have laced their panels with channels filled with an inert gas. Applying a voltage to a channel makes the gas conduct current and so alters the orientation of the molecules just as transistors do. But building these displays is far easier than etching transistors, says Thomas S. Buzak, the Tektronix scientist who invented the technique.

Japanese Spending on Active-Matrix Liquid-Crystal Displays

MILLIONS OF DOLLARS, 1990–1992

700	SHARP (through 1993)
560	SANYO
350	MATSUSHITA
210	HITACHI
140	HOSHIDEN
140	TOSHIBA (with IBM)
70	MITSUBISHI
70	NEC

SOURCE: Nikkei Sangyo Newspaper Survey

"I don't think there is a better display technology in the world," declares Tom Long, vice president of the company. But the prospect of raising more than $100 million for a factory and the uncertainty of earning an adequate return on that investment remain daunting. "Given the capital markets, we will probably have no choice but to license the technology to Japan or Korea," he adds. "It is a sad, sad fact, but I think that's what it will come down to."

Similarly, Texas Instruments "is not in the display business," nor does it want to be, declares L. Eugene Matthews, who directs the computer systems laboratory. On the other hand, prowess in lithography has led TI workers to a novel approach for making projection displays. They are etching tiny mirrors—less than 20 microns wide—in silicon substrates. By changing the tilt of individual mirrors, they can selectively reflect incoming light in different directions and so alter the brightness of the final image.

These mirrors are making their commercial debut in the company's newest generation of high-resolution airline-ticket printers, Matthews explains. For printers, TI needs to cut only a single row of mirrors in silicon. Displays, however, demand arrays of mirrors, say, 2,000 across and 1,000 deep, on a single chip. Fabricating such arrays is still essentially an experiment. "Can we pattern all those mirrors? And build these with good yield?" Matthews asks.

Those are the research questions. On the commercial front, TI may eventually sell modules—chips with arrays of mirrors and associated electronics. Matthews hopes display manufacturers will be interested.

All of these efforts have at least one other common feature: the light in the system is generated from a separate source, typically behind the panel. So-called emissive systems, in which the elements themselves emit light, are often brighter. For instance, Photonics Imaging in Northwood, Ohio, recently introduced a full-color plasma display that measures 17 inches in diagonal and can display video images that are as bright as those on a conventional television, says Peter S. Friedman, president of Photonics. Plasma displays operate by exciting a gas. The ultraviolet light from the plasma stimulates a phosphor coating on the screen. Friedman believes the company is ready to try small-volume manufacturing. "The significant hurdle is that we have to raise the money to do it."

Although their technologies are promising, both big and small U.S. firms are stumbling on that investment hurdle. There is, moreover, little the government can do. "Manufacturing is a private sector issue, not a government issue," says Craig Fields, chief executive officer of the Microelectronics and Computer Corporation in Austin, Tex. "The problem is with investment. Companies can spend millions of dollars, face fierce competition and lose money the whole time. That's not an investment opportunity being sought out in the U.S."

Around the country, companies are talking with one another about the possibilities of joint manufacturing ventures to share the cost and risks. "To date, none have jelled," Fields says. "I hope one does." —*Elizabeth Corcoran*

Lighten Up

Memory cards are key to the truly portable computer

Why does the average laptop computer weigh in at a hefty seven pounds? Because it takes three pounds of batteries to run the mechanical disk drive that allows the computer to store data after it is switched off. There may be an alternative. Chip makers and computer makers are hard bent on developing a form of permanent memory free of the drawbacks of magnetic recording media.

The solution may be packages of silicon chips known as memory cards. About the size of a credit card but a bit thicker, memory cards store programs and data. They are smaller, faster, lighter and more rugged than the mechanical disk drives in today's laptop and desktop computers. Data packed onto the chips can be quickly retrieved with little expenditure of energy. With new hardware and software design standards in place, memory cards compatible with virtually any personal computer promise a quantum jump in portability over the next few years.

Just one fully functional portable personal computer on the U.S. market bears a memory card so far—the one-pound Poqet PC manufactured by Poqet Computer Corporation in Santa Clara, Calif. The $995 videocassette-size IBM-compatible, introduced in September 1989, can run for up to 100 hours on two AA alkaline batteries.

Some 10 to 20 card-based portable computers are expected to debut this November at Comdex, the annual computer industry trade show in Las Vegas. Most industry observers anticipate that "palmtops" like Poqet will get bigger, becoming more like the notebook-size computer Fujitsu has begun shipping in Japan; the FMR-Card computer measures 8 1/2 by 11 by 1 inch. Powered by just two batteries, it contains two memory cards and weighs just two pounds. "By 1992 we expect every notebook will have a slot, and users will have a growing availability of memory and network cards," predicts William V. Ringer, product line manager for Intel network cards. The Santa Clara company is one of the world's leading suppliers of computer components.

As the cards store and process ever more information, they promise to expand the applications of small personal computers in ways not yet commonly thought of or widely used. Aside from spreadsheets and word-processing programs, card developers envision plug-in modules containing road maps, restaurant listings, even newspapers. Custom memory cards have already found their way into industrial applications such as aircraft flight testing and inventory control. A few years ago they peeped into popular use for music synthesizers, video-game cartridges, spell checkers and personal organizers like Sharp's Wizard and Casio's Boss.

Memory cards are still relatively expensive at about $350 per megabyte of capacity, but prices are dropping and density is increasing fast. When Poqet decided to bet on memory cards in 1988 the largest available card held

Reading Books Byte by Byte

First came the Sony Walkman portable tape player. Then, when compact discs swept into the music market, the Japanese electronics giant rolled out the Discman. Now the company's latest bid in consumer electronics provides data instead of decibels. The Data Discman can pack about 200,000 pages of text on the same shiny discs that play singles of rock and Bach.

So far the product is on the market only in Japan, but there it is doing very well indeed. In its first six months of sales, consumers have snapped up about 100,000 Data Discman players, rivaling sales of some of Sony's best-selling domestic products, such as its passport-size eight-millimeter video camera. Buoyed by consumers' enthusiasm, Sony may try to introduce the Data Discman in the U.S. and Europe in time for Christmas shoppers, according to some market analysts in Japan.

The idea for the Data Discman was born three years ago, when Sony engineers noticed young employees were delighted by pocket-size electronic notebooks. "We wanted to create an intelligent product," says Jun Tanaka, marketing director for Sony's electronic publishing project in Japan.

The result was a modified version of Sony's portable compact disc player. The Data Discman reads information from eight-centimeter, read-only memory compact discs (CD-ROMs), which can store 200 megabytes of data. Users tap in queries on a small keyboard, then read the information on an attached liquid-crystal display screen.

Sony is avoiding even the hint of an association with personal computers, however, by ensuring that the new device cannot be hooked up to a computer.

"We're saying this isn't a computer, even though it uses computer technology," insists Jeffrey C. Marshall, a member of the Sony team developing the Data Discman. "It should be easier and friendlier to use than a computer."

No one is likely to read a major literary work from start to finish on a Discman. A user can, however, exploit the product's search capabilities to dip into reference sources such as dictionaries, phrase books as well as medical and tourist guides for handy facts.

To assure that software was available when the hardware hit the streets, the company lined up a string of publishers who pledged to support the Sony-sponsored standard, which was called EB (for Electronic Book). "The software is the key to defining how the information on the disc can be used and how it appears to the user," Marshall says. Data Discman players store a single information-retrieval program, which offers up to six searching strategies for finding information on discs. Individual disks need comply only with the Electronic Book standard.

As a result, "every Electronic Book acts the same and has the same kind of structure. You find information the same way, and you read it off the screen the same way," Marshall says. These electronic books are organized like conventional texts: each one has a table of con

512 kilobytes and cost $550. "The decision not to use rotating media was a difficult one," recalls Steven D. Cox, manager of the firm's personal-computer enhancement group. "But for the market we were going after, size and portability were the two main factors."

That market could be very lucrative indeed, says John Reimer, vice president of marketing at SunDisk in Santa Clara and president of the Personal Computer Memory Card International Association (PCMCIA). The worldwide appetite for portable computers is expected to go from 2.9 million units in 1991 to 10 million by 1993. Poqet predicts that palmtops will make up 31 percent of the market; laptops, 28 percent; and notebooks, 41 percent.

Reimer was instrumental in overcoming a major hurdle to developing the cards—namely, standardization. In June of 1989 he founded PCMCIA because "everybody was doing it and no two cards looked alike." A year later the association had more than 100 members. "People saw the light," he says.

Today a triumvirate of semiconductor, software and PC manufacturers is rallying to put memory cards into everyone's mind. Among those companies pursuing the technology are Apple, Chips and Technologies, Databook, Du Pont, Epson, Fujitsu, GRiD, Hitachi, IBM, Intel, Kodak, Lotus, Maxxell, Microsoft, Motorola, NEC, OKI Semiconductor, Phoenix, Polaroid, Samsung, Sharp, Texas Instruments and Toshiba. Strategic alliances are rife.

Under PCMCIA's standard, the plugs on memory cards will all have 68 pins. Each pin corresponds to a discrete data-storage signal. One has been set aside for future uses, such as peripheral functions like modems. Not coincidentally, the pin standard coincides with one put forth by the Japan Electronic Industry Development Association (JEIDA) in 1985. U.S. manufacturers wanted agreement so that any machine that uses the common MS-DOS operating system (IBM PCs, for example) could use any card.

Just as there are different functions being developed for memory and input/output cards, semiconductor manufacturers are employing a number of different methods of data storage. The memory chips used in most computers, known as DRAM for dynamic random-access memory, are not suitable for permanent storage, because they do not retain data if electric power is shut off.

The up-and-coming data-storage system many believe will be best for memory cards is known as Flash memory. It does not need a battery to back it up, and it can be reprogrammed electrically. "The real hinge factor is whether semiconductor manufacturers worldwide will turn from making DRAM to making Flash. We think absolutely yes," declares Jim Weisenstein, director of Intel's Flash card systems group.

The drawback to Flash is that it must be erased in sectors. Instead of changing a message like "Hello, Sarah, how are you doing" to "Hello, Stan..." by simply changing the name, Flash will save an entirely new document, eating up valuable storage space. Some companies are trying to overcome this by erasing smaller and smaller blocks; others will rely on software to move around and erase data while the user is working on another area of the card.

"The goal is to make products so the end user doesn't have to know the foggiest about what's going on inside," says Daniel Sternglass, chairman of the PCMCIA software committee and founder of Databook in Ithaca, N.Y., which makes devices that allow data to be exchanged between cards and conventional desktop computers.

The different types of chips require that data be stored in different ways, he notes. To overcome this obstacle to interchangeability, PCMCIA and JEIDA agreed to standardize a "Metaformat"—a software header that describes for any conforming machine what is on a disk and how it is organized.

"The problem is compatibility of media between older and newer versions of MS-DOS," explains Michael Dryfoos, development leader for MS-DOS at Microsoft Corporation. "Can you see the files, how do you get to information, how do you map the card's memory in an interesting and useful way? It's a hard problem to solve in a way that's clean and friendly."

There are still snags to be smoothed, but the field is sure to undergo evolution as memory-card technology improves. Aggressive, quick companies will do well in the short run, until manufacturing costs inevitably become a more stringent survival criterion. "We'd like to say these big Japanese semiconductor manufacturers weren't aiming their guns right at us, but they are," confesses Intel's Weisenstein. The U.S. should not be counted out yet, however. There are a number of little companies with impressive technologies, and plenty of incentive to lighten up and carry on. —*Deborah Erickson*

Replacing Disk Drives with Memory Chips

- **SRAM:** Static random-access memory requires electric power to retain data, but requirements are far less than for conventional DRAM (dynamic random-access memory) chips. SRAM cards currently cost $400 to $500 per megabyte.
- **Pseudo-Static RAM:** These chips are a compromise between DRAM and SRAM; they include extra circuitry to reduce power consumption. They cost 30 percent less than conventional SRAM cards.
- **EEPROM:** These electrically erasable programmable read-only memories can store data indefinitely without backup power. Their cost, however, is prohibitive—about three to four times the price of Flash.
- **Flash RAM:** Developed by Intel, these chips are technically known as electrically, selectively erasable read-only memory. Like EEPROM, they do not require backup power. Memory cards holding four megabytes are available for about $1,000.

tents, chapters and an index. "The key is standardization," Marshall emphasizes.

At present, some 63 Japanese publishers, electronics manufacturers and other companies are supporting the EB standard. More than 30 EB publications are on store shelves. Most of these are reference books targeted at middle-aged businessmen, such as a collection of Japanese baseball statistics, and *Kojien*, the standard Japanese dictionary. (At least one publisher has put out a disc of recipes, however.) The EB discs range from $25 to $155 apiece; Data Discman players sell for almost $450.

The company also hopes to copy its strategy—and success—in the U.S. and in Europe. In February, Sony America unveiled a new electronic publishing subsidiary. Around the same time in Japan, Sony's EB standard won the backing of Matsushita, Canon and six other major Japanese electronics producers. The move seems aimed at avoiding a bitter conflict over standards, such as the one that took place between Sony's Beta videotape format and VHS.

Despite the flurry of activity, company officials are remaining tight-lipped about whether they have found any U.S. or European publishers willing to tow the Sony EB standard line. If Sony can pull it off, futurist visions of paperlessness may be only a couple of chapters away. —*Tom Koppel, Tokyo*

Clot Spotter

Does a test for fibrinogen predict risk of heart disease?

As if blood pressure, cholesterol, cigarette smoking and weight were not enough, here comes another potential risk factor for heart disease: fibrinogen. This blood-clotting factor circulates through the body on the lookout for any lesion in need of repair with a clot or scab. But lately the protein has been showing up in studies that link high levels of it to heart disease and stroke.

"There is a great deal of interest in fibrinogen as a risk factor," acknowledges John C. Hoak, director of hematology at the National Heart, Lung and Blood Institute in Bethesda, Md. But, he adds, the plethora of studies making the correlation may be raising unwarranted suspicions. It is still undetermined whether elevated fibrinogen levels actually cause heart disease or are simply associated bystanders.

The scientific uncertainty has not deterred Henry L. Nordhoff, president and chief executive officer of American Biogenetic Sciences (ABS), a tiny biotechnology company on the campus of the University of Notre Dame in Indiana. "Fibrinogen should be on a routine physical," he asserts, citing independent studies done in the U.S., U.K. and Sweden, which conclude fibrinogen is a risk factor that should be a target of routine screening—and perhaps therapy.

Nordhoff happens to be developing a rapid, highly specific test for fibrinogen. He hopes a major diagnostic company will license the test, trade named Cadkit, which has been designed to run on the high-volume automatic machines found in clinical laboratories. A bedside version that works on whole blood is also under development. It could be used to monitor heart-attack patients receiving or being considered for clot-dissolving therapy.

One of the studies that puts the onus on fibrinogen is the same one that fingered cholesterol. The Framingham study, which since 1948 has tracked the heart health of residents in the Massachusetts town, measured fibrinogen levels just once, in 1968—among 1,315 people free of cardiovascular illness. It has since found that "if we follow people long enough, those with higher fibrinogen values are at greater risk for atherosclerotic problems," says William B. Kannel, lead researcher for the Framingham study.

The classic way of measuring fibrinogen is with functional assays that add thrombin to a blood plasma sample. Thrombin converts fibrinogen to fibrin, the main component of insoluble clots. The faster a clot is formed, the more fibrinogen is deemed present. A newer testing approach uses monoclonal antibodies, but the clotting process releases breakdown products that can confuse these immunochemical assays.

ABS is betting on a new production technique to manufacture monoclonal antibodies, which the company claims are much more specific to fibrinogen. Instead of isolating antibody-producing cells from normal laboratory mice, ABS creates its antibodies in mice that have been immunologically isolated for generations.

Because these mice, which the company has licensed from Notre Dame's Lobund Laboratory, have developed very few antibodies on their own, the response is highly specific when they are injected with a specific antigen—in this case, human fibrinogen. ABS says that one in 48 spleen cells from these mice produced antibodies to fibrinogen, compared with one in 60,000 cells from normal mice. The cells are then cultured to provide commercial quantities of the antibody. "Antibodies made this way seem better able to 'see' the difference between fibrinogen and fibrin," explains Paul E. Gargan, director of protein biochemistry at ABS.

Given an accurate test, patients with high fibrinogen levels could minimize other risk factors, such as smoking, Nordhoff says. They might also be candidates to receive fibrinogen-reducing drugs, he suggests. A number of medications on the market reduce fibrinogen, although they were not developed to do that specifically, including beta blockers for high blood pressure.

So begin administering fibrinogen-lowering drugs? "God, no," says Greg Vercellotti, an associate professor of medicine and hematology at the University of Minnesota and an expert in clotting. Fibrinogen is an acute-phase reactant that increases with inflammation, he says, "so you'd wonder if the patient had a hidden cancer or infection."

Framingham's Kannel acknowledges that no one really knows how elevated fibrinogen levels could cause heart disease, but he suggests possible mechanisms: with more fibrinogen around, it is more likely a clot will form; higher levels also could be a measure of the activity of lesions and the body's efforts to repair them. "The significance of rising or falling levels of fibrinogen is a continuing saga in research," Kannel notes. ABS would like to write itself into the story. —*Deborah Erickson*

Starting from Scratch

Gordon Bell is clearly a man after the hearts—and businesses—of aspiring entrepreneurs. In the 1970s he led the design team that built the VAX superminicomputer at Digital Equipment Corporation in Maynard, Mass. That work helped Digital blossom from a midsize company into one that IBM had to reckon with. Since then, Bell has had a hand in some 20 start-ups, most of which have tried, with varying degrees of success, to build computer hardware.

Now Bell is trying a different kind of start-up, one that aims to guide other start-ups through infancy. He has developed what he calls a diagnostic, a collection of questions he thinks could help fledgling entrepreneurs. Bell's approach is outlined in his recent book, *High-Tech Ventures* (Addison-Wesley), which reads a bit like a Dr. Spock guide for young companies. Among his questions: Will the company need fewer than three technical breakthroughs to make its product? Do the designers have a manufacturing strategy in mind? Will the product still be trendy when the company seeks additional funding?

Yet building a successful start-up is still more of an art than a science. Some consultants may disagree with Bell's priorities, even on elements as standard as a business plan. Bell argues that such a plan should serve as a dynamic blueprint for the company. In contrast, Edward B. Roberts, a professor at the Sloan School of Management at the Massachusetts Institute of Technology, suggests in his recent offering, *Entrepreneurs in High Technology* (Oxford University Press), that his studies reveal few relationships between high-tech start-ups, their initial plans and their later performance.

So far, Bell says, Japanese managers have been more intrigued by his approach than have U.S. venture capitalists. All Bell hopes is that his pointers will steer some entrepreneurs away from the pitfalls he has encountered. As he wryly observes: "In hardware engineering, Ohm's law and Maxwell's equations pale in importance and influence next to Murphy's law." —*Elizabeth Corcoran*

Selling Cells

Is a kidney cancer treatment a therapy or an experiment?

Malignant kidney cancer will probably kill more than 10,000 Americans this year, and 24,000 new cases may be diagnosed. Eugene P. Schonfeld of the National Kidney Cancer Association is blunt about the treatment options for patients whose disease has spread: "Chemotherapy is virtually worthless. It works for fewer than 10 percent of the patients," he says. "You want something 'proven'? Something the Food and Drug Administration approved? There really isn't anything."

There is, however, something the FDA has not approved. Cellcor Therapies in Newton, Mass., is rushing to market a $22,000 treatment called autolymphocyte therapy (ALT). Because ALT is a therapy and not a new drug, the FDA does not control it: medical practice does. The company, which opened a treatment center at the New England Baptist Hospital in Boston in March and plans several more around the country over the next few months, says ALT can more than double the average survival time of some patients without subjecting them to harsh side effects.

Some observers worry that Cellcor is being unduly hasty in bringing ALT to market. Although no one suggests that ALT is dangerous, there are concerns that too little clinical research exists in support of its effectiveness. "It's a perfectly rational approach to explore," remarks Andrew F. Dorr, a medical oncologist at the National Cancer Institute. "But I wouldn't be comfortable sending my mother for this treatment right now."

ALT is an immunotherapy, a procedure in which physicians try to combat tumors by enhancing the patient's immune system. Most immunotherapies involve drugs such as interferon and interleukins—compounds called cytokines that the body naturally secretes to boost immune function. But therapeutic doses of these drugs can make patients profoundly ill, Schonfeld says, and it is not yet clear that they can significantly lengthen most patients' lives.

The key to ALT, explains Michael E. Osband, one of its developers and a cofounder of Cellcor, is that it relies on whole cells, not drugs. Before the treatment begins, physicians extract lymphocytes, or white blood cells, from the patient. With monoclonal antibodies, Cellcor's technicians stimulate the lymphocytes to secrete large quantities of cytokines. Three days later, the technicians discard the cells but save the cytokine-rich medium, which they divide into six portions and store.

Once a month for the next six months, the patient returns to the center and donates another batch of lymphocytes. These cells are added to one portion of the cytokine mixture. After a week of incubation, the activated lymphocytes are reinfused into the patient to attack tumor cells. Because the patient receives only his own cells, activated by his own cytokines, the therapy does not require FDA approval.

The largest study of ALT appeared in the *Lancet* in April 1990—less than a year before the Boston center opened. That study demonstrated that patients who underwent the procedure survived on average 22 months—about two and a half times as long as those receiving a chemical therapy. Some cancer specialists are concerned, however, that it is the only randomized trial ever published and that it involved only 90 patients. In their minds, ALT is still a treatment in its experimental stages. "With so few patients in a study, the chances of the treatments being different through chance occurrence are not insignificant," Dorr says.

Neither Cellcor nor its skeptics would like to see a repeat of the Biotherapeutics controversy. In 1984 that Tennessee-based biotechnology start-up offered cancer patients the opportunity to participate in research programs and receive experimental treatments—for a fee. After critics savaged it, Biotherapeutics retreated from that business in 1989.

Osband and Richard R. D'Antoni, Cellcor's president, argue that comparisons between Cellcor and Biotherapeutics are false. "We have never charged patients for therapy that's experimental," Osband says. He also points out that Cellcor is selective in admitting patients to the commercial program. "One of the things that we're proud of is that we're not treating everybody with cancer or even everybody with metastatic kidney cancer." Only patients who fit the profile of those helped in the randomized trials, he says, are allowed to participate, and all of them receive exactly the same treatment.

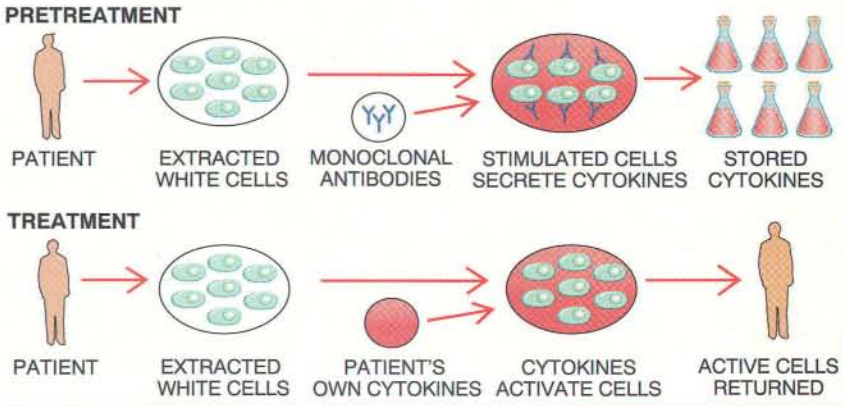
More randomized trials would be persuasive, Osband agrees, "but we don't think we can ethically do the studies in which we deny this treatment from some people. I have too many data that say this works to withhold it."

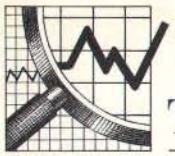
Cold, hard cash may be the final determinant of whether the marketplace sees any treatment as sufficiently proved. If so, ALT seems to be making headway. D'Antoni says that about two thirds of the commercial insurance carriers in the country, including six Blue Cross/Blue Shield plans, are reimbursing for ALT for metastatic kidney cancer. Dorr reports that he receives many calls from insurance companies asking whether ALT and other new treatments for various diseases are appropriate therapies. "I don't know that the Department of Health and Human Services has ever been charged with the duty of defining what's experimental," he says, "but that's what we're being asked to do."

As gene therapy, in vitro fertilization, bone marrow transplants and other procedures involving living cells become more commonplace, the line between experimental and accepted treatments may continue to blur. D'Antoni says that "we think the FDA will eventually be regulating cell-mediated therapies." Whether or not the FDA does so, some regulations or standards may be needed to ensure that patients are getting timely benefits from valuable new therapies without falling prey to medical biotechnology's sales pitches.

—John Rennie

How Autolymphocyte Therapy Works





Sorting Out Chaos on Wall Street

For two decades, the random-walk hypothesis has been the major economic explanation for price fluctuations in financial markets such as the New York Stock Exchange. According to economists, it should be impossible to spot patterns in the streams of data that rush across traders' Quotron screens every day and to make money by betting on the direction of those trends. Traders who seem to make above-average profits by such speculating, they say, are simply lucky (or have inside information).

Yet even economists do not deny that some traders are astonishingly lucky. And increasing numbers of financial firms seem willing to spend money and time trying to boost their "luck" by developing proprietary computer software to look for patterns in market data. "There is no rigorous proof of the random-walk theory," points out John Geanakoplos, an economist at Yale University. As a result, earlier this year the Santa Fe Institute in New Mexico convened a curious assembly—some 70 economists, physicists and financial brokers—to talk about whether stock-market patterns exist.

To make money by juggling the numbers rubs economists the wrong way on several accounts. To begin, all market players see the same basic data—the ups and downs of a share of IBM, for instance. If new information suggests that a stock is undervalued, enough people will place buy orders that the price will swiftly rise and the stock will be fairly valued again.

Looking for trends in a hodgepodge of more complex data divorces the numbers from real-world economic events. Brokers would do as well to scrutinize midtown Manhattan traffic patterns, the economists suggest. And the final, telling anecdotal evidence: economists have had dismal luck nailing down successful trading rules.

Maybe so. But what if the economists are simply using outdated statistical techniques? At the Santa Fe meeting, physicists and computer scientists came equipped with toolboxes of sophisticated algorithms to ferret out nonlinear patterns in physical data. Researchers once believed that turbulence in liquids was simply random, points out Doyne Farmer, a researcher in complex systems at Los Alamos National Laboratory. Farmer was among those who demonstrated that such systems actually show different degrees of order.

Neural networks might also help, suggested Richard G. Palmer, a physicist at Duke University. These software systems construct internal models of the world by assigning heavier weights to the connections associated with the input data received most frequently. Sometimes such models fail to recognize a clear pattern, Palmer concedes. Still, they can often highlight dynamic and unexpected trends in data. In contrast, in traditional economic models people only learn expected results or patterns. (For instance, trader A learns what trader B already knows.) Both neural networks and genetic algorithms, in which successful rules breed ever more successful rules, could help incorporate a more realistic model of learning, Palmer observes.

The traders, moreover, argue that the economists have not found patterns in data, because they do not work hard enough at "scrubbing," or cleaning up noise in financial data. Even the most clever neural networks will not find patterns in messy data, they say. "Your [nonlinear] models are like Lamborghinis or Ferraris," David J. Hirschfeld, a director of commodity research for Tudor Investment in New York City, told the group. "But you're putting apple juice in the engine," while "we're driving BMWs but with high-quality fuel."

"Group psychology may have a more potent effect on the markets than economists have typically believed."

Tudor has been undeniably successful with its trades; Hirschfeld reports that the group has racked up a 94 percent average return for the past six years, operating with about $750 million in the U.S. "I'd submit that I could probably outtrade every [sophisticated nonlinear algorithm] using just three or four indicators because of the way I transform data," he declares.

Nevertheless, brokers are often hard-pressed to explain precisely why they capitalize on specific trends. W. Edwin Bosarge, Jr., who runs the Houston-based firm, Frontier Financial, claims to have developed algorithms that analyze patterns based on equations from Newtonian physics. "Why should Newtonian physics work? I haven't the slightest idea," he says. "But I know it does, and so does quantum mechanics."

Even Tudor's star trader, firm chairman Paul Tudor Jones II, has trouble explaining why he finds some patterns in market data interesting. "He will point to a relationship in data, and we'll look for other relationships," Hirschfeld says. But, he adds, "it rarely works out that what he did is successful for the exact reasons he proposed."

If there are recognizable albeit faint patterns in market data, what could cause them? Answering this question could have implications for economists that extend beyond the possibility of increasing their bank accounts. For instance, suppose a handful of clever traders can interpret the economy. When these few execute their trades, they may leave telltale tracks in the market, says Geanakoplos, who co-chaired the Santa Fe meeting. Others need not understand the economy—just follow in the smart traders' wake.

On the other hand, group psychology may have a more potent effect on the markets than economists have typically believed. Traders may go through predictable mood swings—say, becoming aggressive buyers after a short run of successes and then more hesitant as they expect their streak to fizzle. If nonlinear analysis techniques show patterns that indicate traders' moods play a significant role in moving markets, then economists will have to scramble to find better ways to incorporate psychology in their models. Finally, subtle underlying patterns may indicate structural relationships in the economy that economists have not yet seen.

Still, before economists completely abandon random walks, they have to devote more energy to examining data for quantitative proof, Geanakoplos points out. But even a cursory calculation of the risks and rewards indicates that it would be a valuable investment of time. —*Elizabeth Corcoran*

A Swift Trip over Rugged Terrain

In the summer of 1990 Irena Watts, director of book conservation at the Bodleian Library in Oxford, England, made a discovery of Brobdingnagian proportions. As Watts repaired the binding of a Jacobean psalter, she noticed that the book cover was made from several layers of paper that had been stiffened with glue. She managed to separate the sheets without destroying their contents. The papers turned out to be a hitherto unknown chapter of Jonathan Swift's masterpiece, Gulliver's Travels. SCIENTIFIC AMERICAN *is proud to publish, for the first time, "Gulliver's Further Adventures on the Flying Island of Laputa."*

In about a month's time I had attained tolerable proficiency in the Laputan language and was able to answer most of the King's questions about the state of mathematics in Europe. As a gracious gesture in return, the King invited me to inspect the Laputan Flying Academy, modeled after the much larger institute that rested upon the firm earth of Balnibarbi Island. I made haste to accept, out of curiosity as much as politeness, for I had heard many tales about the academy and its natural philosophers.

I was first introduced to a kind inventor who had for 16 years been attempting to build a grandfather clock using a double pendulum—the better, so the inventor assured me, to tell the time. His original intention had been to suspend the second pendulum from the end of the first, in as simple a manner as possible. Upon realizing that certain subtleties of his theory could not be borne out in practice, the inventor had perforce added a spring here, a counterbalancing weight there, so that during the 16-year period of development the machine's complexity had greatly exceeded that originally envisioned [*editor's note: see illustration on next page*]. Enquiring how accurately the clock performed, I was told that it was correct twice each day.

The clockmaker became quite friendly when I complimented him on this excellent standard of performance, and he showed me the many mathematical calculations that determined the machine's design. Although I am unable to recall them all, one has remained stubbornly in my memory. A fact central to the successful operation of the clock was that its pendulums should on occasion be at rest, so that the combined action of the springs should come to perfect balance. The simplicity of the original design, he informed me, made such calculations both elegant and straightforward: the pair of pendulums could balance in exactly four positions.

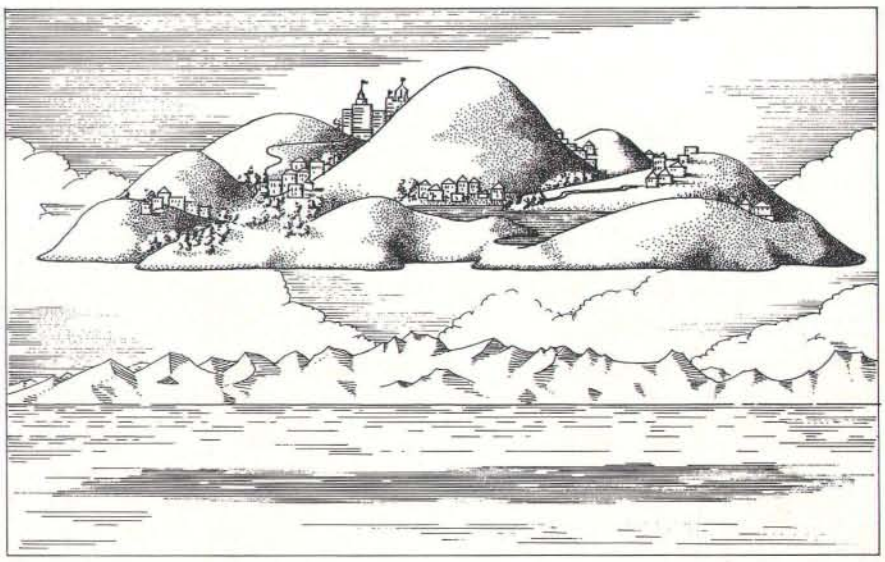
I understood at once that if both pendulums were to hang vertically downward, then they would remain forever at rest in that position. I begged, however, to dispute the existence of three other such positions. With the expenditure of much effort, the clockmaker led me to understand that a second such position was possible, with both pendulums pointing vertically upward.

I admitted that such a configuration had not occurred to me, and upon being informed that in principle such an arrangement could be made to balance, I observed that although in theory a monk can balance an eel on the end of his nose, such behavior is seldom seen in the fish markets or the monasteries. But the clockmaker persisted, describing the arrangement as a *thelmin frole,* which I translate as "unstable equilibrium." At the time I believed he said that the arrangement was "fiendishly unlikely," a sentiment with which I had hastened to agree. Having understood this, I was quick to deduce the two remaining equilibrium states, in each of which one pendulum is balanced vertically upward and the other hangs vertically downward.

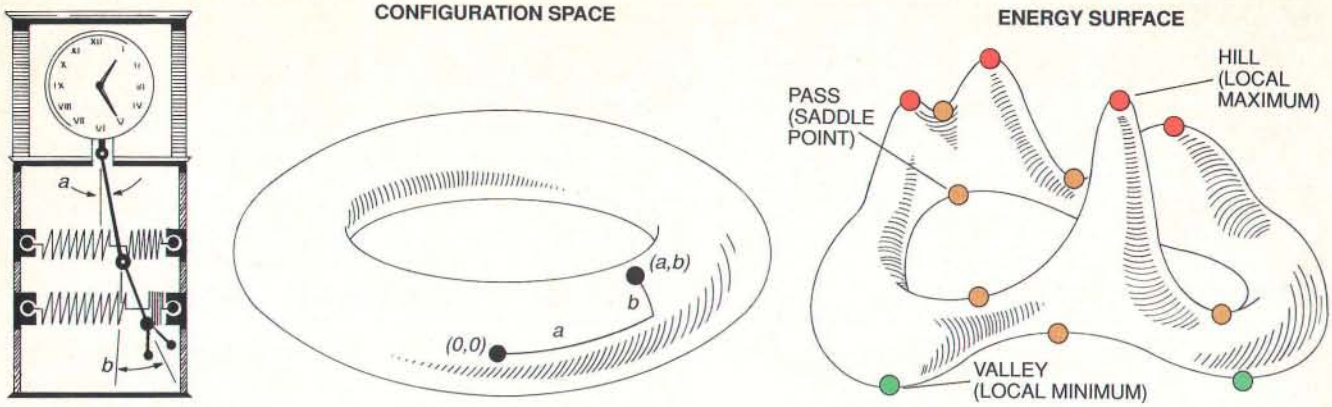
The inventor lamented that the sole obstacle to completion of the project was to establish the existence of four or more comparable configurations from the actual apparatus, springs and all. The precise positions were unimportant: all that was required was that they should exist. But the mathematics was proving impenetrable, and he despaired of ever finding an answer. At that point I was rescued by the King's messenger, who bade me repair to the kitchens for a meal of ellipsoid pudding and conical beef before continuing my visit of the academy.

After the meal I was shown something of which the King was very proud, a geodetic survey of the entire island. The Surveyor Royal was a ruddy-visaged person of huge girth, who always carried a plumb line and bob as a badge of office. His task, he told me, was to catalogue every hill, valley and pass upon the island.

I enquired as to the precise definition of these terms, wishing to apprehend the exact nature of the enterprise. Was an anthill, for example, accounted for as a hill? He said proudly that it was. A hill was any prominence whose height



ISLAND OF LAPUTA has been blessed with 1,267 hills, 1,506 valleys and 1,944 passes, according to the Surveyor Royal. Can you prove that he miscounted?



MECHANISM of the great-grandfather clock **(left)** ***is based on two pendulums. The angles of the pendulums correspond to points on a surface*** **(center),** ***which is known as the configuration space of the clock. If each point of the surface is displaced to a height equal to the energy of the corresponding configuration, the energy surface*** **(right)** ***is formed.***

exceeded that of its immediate vicinity; it was a flam, or "local maximum." A valley, correspondingly, was a flim, or "local minimum." The Surveyor Royal explained that a pass was technically a place that had a local maximum in one direction but a local minimum in another, akin to the saddle of a horse. Indeed, such a place is known as "a saddle point" in the great European academies, but the Laputan term for a pass is, of course, "flimflam."

The enumeration of these features, he explained, was performed to the utmost precision. By this reckoning, he said proudly, there were exactly 1,267 hills and 1,506 valleys in Laputa. At this juncture I interposed that there must therefore be 2,771 passes.

He claimed that the Royal Geodetic Survey had mapped precisely 1,944. I replied some must have been omitted, for there is a relation between the number of hills, H, valleys, V, and passes, P. The relation is $H+V-P=2$. The proof of it, I informed the Surveyor Royal, is both general and conclusive [*editor's note: see box on opposite page*].

I informed the Surveyor Royal that he had miscounted the number of passes by 827, always presuming that he had made an accurate census of the number of hills and valleys. In support of the official count, the Surveyor Royal paraded before me his subordinates by the dozen to swear to the accuracy of their observations. But piece by piece, discrepancies began to appear, and shortly the Surveyor Royal announced that because of a small oversight the number of passes must be augmented slightly, to become 2,772, the number of hills and valleys remaining unchanged.

I applauded his diligence but ventured to remark that there remained a discrepancy of one. Either he had overestimated the number of passes or had overlooked a hill or a valley. He objected that a hill, by its nature, cannot be overlooked, but he also admitted that, by the same token, virtually any valley might be overlooked.

Then he became greatly excited and took me before yet another member of the academy, a historian. I confess that less likely a personage for such an office I had never observed, for he could scarce remember his own name from one second to the next. By dint of great application on his part, however, and even greater patience on mine, a semblance of a tale began to unfold.

I was not, it seemed, the first European visitor to Laputa. One Captain Kidd, a pirate by trade, was rumored to have buried a treasure somewhere on the island, "at the bottom of its deepest valley." Despite repeated searches, no such treasure had been found, but perhaps my contention that a valley might have been overlooked would resolve the mystery. It was unfortunate that my method did not reveal the precise location of the missing valley.

Upon thinking the matter over, however, I realized that the "deepest valley" on Laputa could only be the very lowest point on the underside of the island, which was smooth and gently rounded, like a dish. Discreet enquiries confirmed that the Royal Geodetic Survey had not included the underside of the island, and the discrepancy was resolved to my own satisfaction. I vowed to inspect the nethermost region of the island for myself.

As a distraction I asked whether any natural stone arches could be found on Laputa. I was told by the Surveyor Royal that, indeed, there were several arches, although he did not know exactly how many. I informed him that the Royal Geodetic Survey must still be in error. My proof that $H+V-P=2$ had assumed the absence of holes in the island. On any surface that, like stone arches, possesses holes, the relation between the numbers H, V and P must be different from the one that I had previously derived. Now $H+V-P=2-2g$, where g is the number of holes.

A truly accurate survey establishing the number of hills, valleys and passes without error would make it possible to deduce the number g of stone arches. For instance, if there were 1,000 hills, 1,000 valleys and 2,020 passes, then the following relation holds:

$$1{,}000+1{,}000-2{,}020=2-2g$$

whence $2g=22$, so that there would exist precisely 11 stone arches. I do not believe that the Surveyor Royal was pleased by these revelations, but he pledged on the spot to repeat the survey in total accuracy.

My mind had never been far from Captain Kidd's treasure. It occurred to me that the pirate might well have buried his treasure in a deep cavern called Flandona Gagnole, which I had been shown earlier in my visit. Within this cavern is a lodestone of prodigious size, sustained upon an axle. By means of the lodestone, the island is made to rise and fall and is conveyed to different parts of the world. Perhaps the treasure was to be found beneath the lodestone, this being as near as practicable to the island's nethermost point. I resolved to dig for it and to this end secured a spade from the King's gardens. Regrettably, before my tunnel had been dug more than a few yards, I was apprehended and imprisoned for alleged sabotage of the Flying Island.

I lay in chains for three days and was then taken before the King, who belabored me mightily for my infringement of Laputan law. At length I prevailed upon him to hear my plea of mitigation, and I revealed my deduction that

the great pirate treasure was to be found on the underside of Laputa. The King expressed his gratitude for my ideas and then sentenced me to four days of hard mental labor.

My first task, as one might expect, was to find a way to recover the buried treasure for the King. I decided it would be best to consult with several members of the Laputan Flying Academy. One philosopher asserted that while my plan to tunnel downward from Flandona Gagnole was sound in principle, it was better carried out by reversing the lodestone, thus causing Laputa to fly upside down. I quickly pointed out that all the island's inhabitants would thereby fall off. My objection was quickly dismissed by another philosopher, who suggested that all people and possessions could be secured by liberally applying a strong glue over the whole island.

Desiring to avoid such sticky business, I decided that the simplest solution would be to lower Laputa to within a few yards of the ground and erect a ladder to investigate its underside. I considered proposing this myself but became apprehensive lest I be sent to try it, only to have the island accidentally lowered upon my head.

Perhaps I could find someone else to relay my plan to the King. The clockmaker seemed an ideal candidate, being well respected by the academy. He agreed to deliver my modest proposal to the King—but only if I helped him prove that a system of two pendulums has at least four equilibrium states, no matter how many springs and counterweights are attached.

We agreed that the possible configurations of the machine are defined by two angles, those of the two pendulums. These two angles naturally correspond to the points on a particular surface, namely, a torus, which I shall call the configuration space of the machine. This torus has one hole, and hence its genus is $g = 1$. Thus, no matter how the torus is arranged in space, the number $H + V - P$ always vanishes.

When a spring is compressed, or a weight raised, it requires the expenditure of considerable amounts of energy. Thus, associated with any configuration of the mechanism of the clock, there is a mathematical quantity, the total energy.

Imagine each point of the toroidal surface displaced to a height equal to the energy of the corresponding configuration, and let this be named the energy surface. Equilibrium states correspond to positions of stationary energy, that is, to hills, valleys and passes on this surface. Thus, the total number of equilibrium states is $H + V + P$. Now, every finite surface must have a highest point, hence at least one hill, and a lowest, hence at least one valley, whence H is at least 1 and V is at least 1. Since $H + V - P = 2(1) - 2 = 0$, it follows that $P = H + V$, and therefore P is at least 2. Finally, the number of equilibrium states $H + V + P$ is at least $1 + 1 + 2 = 4$, which is the result required by the clockmaker.

My method does not specify where these equilibrium states are, but it successfully provided the desired lower bound for their number. I remark that the conclusion is quite independent of springs, ropes, weights and other embellishments: it depends solely upon the number of holes in the configuration space. I hastened to explain my reasoning to the clockmaker, and after considerable debate, my proof was adjudged sound, albeit outlandish.

The clockmaker kept his word and proposed to the King that the Flying Island be lowered toward the ground to conduct a proper search for the pirate treasure. The King agreed and gave the order. Before the end of the day, Laputa was hovering a few yards above the solid ground of Balnibarbi Island.

As a rope ladder was readied for a landing party, the Surveyor Royal begged an audience to inform the King that the Second Royal Geodetic Survey had been completed. The numbers, he said, scowling darkly in my direction, were 1,893 hills, 1,942 valleys and 3,816 passes. The King seemed pleased to hear the news. By my reckoning, however, the number of g of stone arches must perforce be given by

$$2 - 2g = 1{,}893 + 1{,}942 - 3{,}816 = 19$$

so that Laputa possessed -8.5 arches.

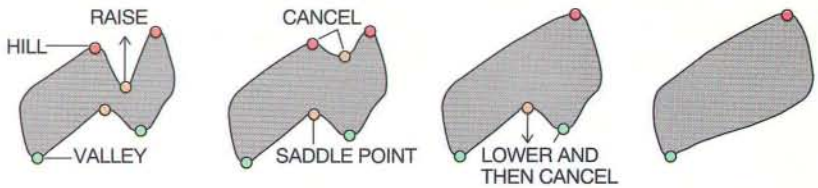
Before the King could become apprised of this fact, I took my leave and sneaked down the rope ladder to the ground. Despite the fear of being crushed, I was tempted to look for the treasure on the underside of the island, but upon hearing a great commotion from above, I departed in haste.

Gulliver's Theorem about Surfaces

Any closed, smooth surface that includes a number H of local maxima, V local minima and P saddle points always satisfies the equation $H + V - P = 2$. To prove this, the best strategy is to deform the surface continually, reducing these numbers in such a way that the expression $H + V - P$ remains unchanged.

The deformation consists of a series of moves. In each move, either a hill or a valley is merged with a neighboring pass, so that both disappear. The process continues until all passes have been eliminated, after which there can remain only one hill and one valley (because if there are two hills, or two valleys, there must be a pass somewhere between them).

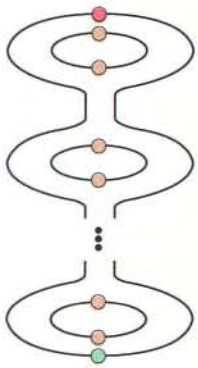
Merging a hill and a pass decreases both H and P by 1, and so $H + V - P$ remains unchanged. Similarly, merging a valley and a pass decreases both V and P by 1, and so $H + V - P$ again remains unchanged. The illustration below shows the effect of a sequence of such moves.



When the deformations are completed, $H = 1$, $V = 1$ and $P = 0$, and so $H + V - P = 1 + 1 - 0$, which then equals 2. Because $H + V - P$ is unchanged throughout the sequence of deformations, the value of $H + V - P$ at the beginning must also be 2.

The proof of $H + V - P = 2$ assumes the surface is topologically equivalent to a sphere. A general surface, however, is equivalent to a sphere with a number g of holes bored through it. The value of g is the genus of the surface. Any closed, smooth surface of genus g must satisfy the equation $H + V - P = 2 - 2g$. The proof is the same as before, but the deformation ends with the surface shown at the right.

Here $H = 1$, $V = 1$ and $P = 2g$ because each hole yields two passes. So $H + V - P = 1 + 1 - 2g$, which then equals $2 - 2g$.



FURTHER READING

SURFACE TOPOLOGY. P. A. Firby and C. F. Gardiner. Chichester, Ellis Horwood, 1982.

BOOK REVIEWS by Philip Morrison

Noon Darkness

THE UNDER-STANDING OF ECLIPSES, by Guy Ottewell. Astronomical Workshop, Furman University, Greenville, SC 29613 (paperbound, $12.95).

Artist and high geometer, Guy Ottewell is poet enough to build a waggish punning title out of that single hyphen. The impetus for this book is plain: on Thursday, July 11, 1991, the solar disk will be blacked out where the eclipse is greatest, about local noon on the Pacific shores of Mexico. The grand rhythms are so nearly in step then that the new moon can drift along with the sun in the sky to hide it for almost seven minutes, a duration unrivaled for 141 years to come.

No other eclipse before 2017 will bring totality to the broad North American mainlands. None so far in human history has darkened daytime skies above so many human witnesses, for this path passes over half a dozen major cities of Mexico, including its smoggy, giant federal capital, Mexico City, home to more than 20 million. On that high plateau it is the season of summer rain and afternoon thundercloud, so it may be that not many will recognize the shadow passage. Fair skies are expected, though, for the four-minute eclipse near dawn across the big island of Hawaii and for the six-minute shadow pass along desert latitudes at the mouth of the Sea of Cortez. The path continues far to the south—the valley of Cali in Colombia has a good chance of clear skies—until the moon's shadow draws back from the earth at sunset far inland in Brazil.

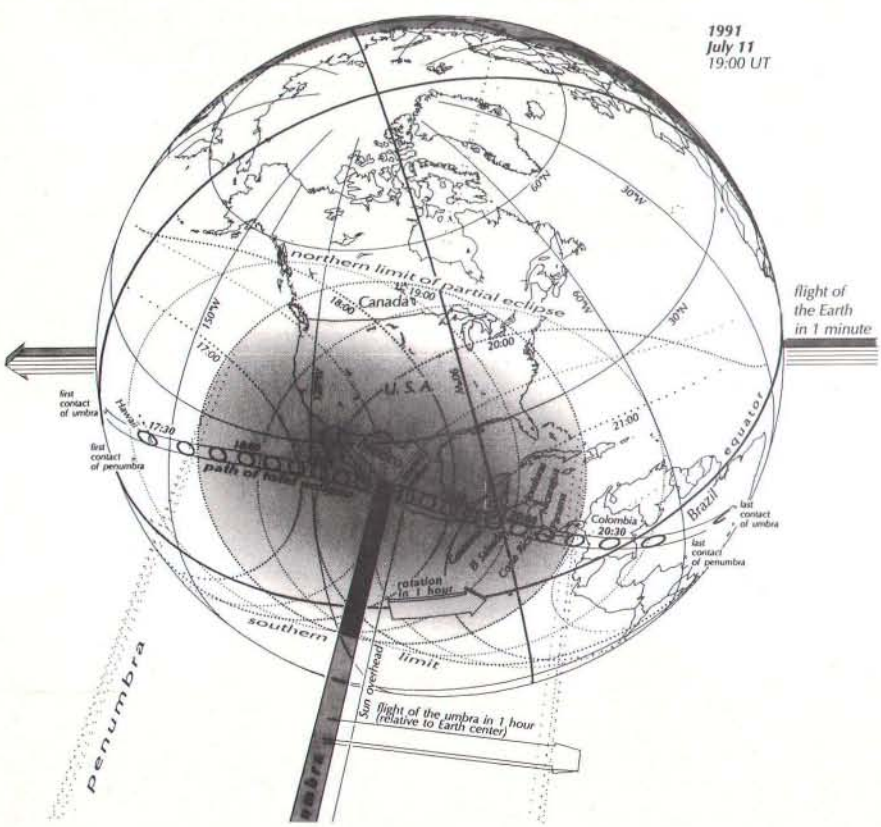
Energetic U.S. amateurs by the hundred thousand may journey to the shadow path. Ottewell's brochure describes well the wonderful sights they can hope to enjoy. During the hour of the partial phase, the blue sky bites off more and more of the disk of the sun. Watch it safely and indirectly on the sun-dappled ground under a leafy tree; not this time, though, for New Englanders at home, but something to be seen for most of the forty-eight.

As totality nears, golden sunlight changes color subtly, in a kind of internal solar sunset. For then we see only rays that have more or less grazed the sun sphere itself on the way out. Brilliant stars and a cluster of planets will shine forth as awesome totality unmasks this midsummer's daytime sky. The moon's notched edge will scintillate, and finally the blackened sun's crimson rim and pearly corona will transfix every witness.

But the originality of this book centers less on under-standing the eclipse than on understanding it. What is most fully diagrammed and explained are the alignments and the size and motion of all of the shadow, near-shadow and antishadow volumes that mark this especially symmetric event among the unceasing run of its varying counterparts, including the eclipses of the moon. The core of the work is a full account of the saros, the ancient half-recognition of an 18-year recurrence that subtly links solar eclipses into finite families, linked strands within a smoothly flowing sequence of aspect and time. What the author calls a "bead-curtain" diagram displays across a dozen pages all eclipse shadows year by year for some two centuries after 1901, in a patterned tour de force of celestial geometry.

Ottewell notes that built-in complementary relations between widely present lunar eclipses and well-localized solar ones may well have helped the less traveled scholars of long ago to pick out what is by no means a simple pattern. At its root lies a remarkable arithmetic truth. At eclipse, three rhythms come in step: the sun-moon alignment month, from one new moon to the next; the nodal month, or the interval between one time when the moon pierces southward through the plane of the earth's orbit and the next; and the monthly cycle of moon-shadow size, the time between one minimum in the earth-moon distance and the next. Eclipses recur in form because 223 alignment times match both 242 nodal returns and 239 distance minima within a few hours out of 18 long years: that coincidence is the saros.

If you can get to the eclipse path, go; this brief book will be well worth carrying. If you stay at home, follow the score here to grasp these harmonious but silent rhythms, and you will have most fitly celebrated the long eclipse.



Total solar eclipse, July 11, 1991

Less is More

THE MACHINE THAT CHANGED THE WORLD, by James P. Womack, Daniel T. Jones and Daniel Roos. Rawson Associates, Macmillan Publishing Company, 1990 ($22.50).

There is on this earth about one motor vehicle for every 10 human heads; automass doubles human biomass. The world's factories deliver every week about one million new wheeled creatures. This readable book is a revealing survey of the largest manufacturing industry, as it was, as it is and as it will be in a decade or two. The view is almost entirely internal: How do they make all those cars? It is a nontechnical but analytic chronicle of auto manufacturing, not of grinding machines nor design blueprints nor salesrooms nor robots, though all of those enter, but rather of the technical and human organizations that have rolled out the ubiquitous machines whose mobility has shaped our lives and our century.

Your car has a metabolic drivetrain held in a more passive body. Since the days of the Ford Model A, that body has resembled a metallic lobster shell more than it has a buggy. The strong and smooth shell is welded from many complex metal shapes. This up-to-date design—little left in it of the horseless carriage—can be built in three distinct ways. Aston Martin in England builds costly and beautiful cars, whose body panels of aluminum are handmade, pounded into form with wooden mallets against smooth die surfaces by skilled craftsmen, men who achieve by eye and hand the graceful forms they seek. That firm has made one car a day for six decades. In Ur of the Chaldees craftsmen produced sculptured forms in sheet copper much in the same way, though hardly so repetitively: the workmanship of risk is very old.

Mass production of big, doubly curved metal panels to be welded into a car body is only decades old. Every maker who produces more than a few cars a day now uses the same scheme, modernized from Henry Ford. Sheet steel is press-cut, "stamped" from the roll into flat pieces. A giant hydraulic press squeezes each cut blank between two expensively and precisely formed hardened dies to impose the three-dimensional shape, say, of a stylish fender. A big die is often half a ton of some exotic alloy. The massive press operates once every five seconds. One year's run stamps out a million fenders, all closely the same. Don't stop the press! Its success is in long, steady runs.

Skill has been replaced by tedium; all risk has receded to design, die making, assembly and sales. The die makers themselves wait until a new model design has been fully specified to begin their computer-controlled cutting. After two years of preparation, the big presses will begin again to stamp, without stopping, from precise new dies the perfected parts for the next model. The economy of scale is obvious; from the factory a couple of thousand cars roll out each day, at a price well over an order of magnitude below Aston Martin's handcrafted elegances. Mass production makes cars by the million, and good ones too.

But a little secret is here disclosed about that "American system" of mass production, born in Ford's Detroit into an auto-hungry world. Down the crowded aisles next to any unceasing assembly line pass many "indirect workers," repairmen, runners, housekeepers, relievers. At many workstations, big bins of parts wait beside trash cans full of what did not fit. At the very end is an enormous work area of cars, finished but with defects, waiting for the expensive craftsmanlike rework. Somewhere outside the windowless plant you can see a large pile of unpainted bodies, amid massive stores of parts not yet unpacked. The primacy of continuity demands costly buffer stocks within every factor of production and promotes the tolerance of frequent error.

There is a third way, opened in Nagoya in the 1950s. The burden of this book is the recognition and description of the new and subtle process. The authors, from M.I.T. and the University of Sussex, call it lean production, successor to Detroit's mass production. In Japan, for example, the die makers work closely with the body designers. Body design work is "lean" as well, its eye on coming changes. The die makers don't wait for precise specs but begin to work when design begins. They use special, flexible cutting tools to form the dies; they are ready to finish dies quickly or to change them once a new model is favored. Even the presses are arranged for quick die replacement.

In the beginning Toyota gave the die-replacement task not to specialists but to production-line workers who would have otherwise been standing idle. Small press runs began; they turned out to mean low stampings inventory, less storage space, fewer specialized workers and more knowledge and interest diffused along the whole line. Errors could be found and corrected quickly before assembly into many cars. Overall costs dropped in spite of many halts in the flow. The same sort of strategems are now found not just in body work but throughout the entire firm. Continuity of flow is no longer sacred; flexibility and freedom from error have replaced the optimized but rigid division of labor.

Lean producers save about half out of each of the factors of production. The scorecards are here, supported by a look at the street. New models in Japan take 0.6 of the engineering time American firms spend. Their strong outside parts suppliers, bound by contract and helped by technical cooperation to deliver to the assembly plant "just in time," do half of all the engineering work, not one seventh of it as in the U.S. After a model change, the return to routine production quality on the line takes six weeks, not 11 months. The buyers get very few lemons; yet new models come out more often at lower cost. The average Japanese "mass market" car remains in production only four years; in the U.S., about eight years. Lean Toyota produces the equivalent of about 29 vehicles per year per employee; massive GM about seven. To be sure, Toyota factories take in from their suppliers about 70 percent of all value added; integrated GM only 30 percent.

An experiment is now being reported in your TV commercials. Will the new Japanese entry into the luxury market reward the lean producers? If luxury car buyers turn out to prefer consistency of design to model change, those producers can offer a wider variety of models at once or else spend more effort on modestly new technology.

Mass production began in autos with Ford, but it is not intrinsically American. Its true home today is clearly in the great European factories, although even there it has lost its Fordian purity. Lean production spreads. American workers, especially in the Japanese-owned Midwestern "transplants," are succeeding. The firm of Ford itself has come a long way. In Amazonian Manaus, Honda employs Brazilian subsistence farmers turned canny and productive assemblers of motorcycles in lean production. Populous China has not gone lean; more autoworkers are employed there than in any other country. Some crowd the big, rigid mass-production plants; some are dispersed among many low-quality craft shops. In China 1.6 million workers produce 0.6 million vehicles a year, while across the Sea of Japan 0.5 million workers produce 13 million good cars each year. The strongest claim of this study is that productivity and success in automobile manufacture at the present time owes more to the organization of the

firm than it does to the national culture.

The book is the report of an expert international field study that sent participants to more than 90 assembly plants worldwide. It has little to say about the great externalities of the auto: fuel, air, roads, safety, community. But its story—far more comprehensive than this account, extending to finance and overseas investment—is compelling. If auto manufacturers are to face market saturation and pressures toward decline with genuine innovation—producing cheap, efficient, collision-avoiding cars, made of composites and fueled with new fuels—it is not hard to judge what production system they will use.

Unruly Reality of Energy

GENERAL ENERGETICS: ENERGY IN THE BIOSPHERE AND CIVILIZATION, by Vaclav Smil. John Wiley & Sons, Inc., 1991 ($69).

Only a work of tightly controlled audacity would dare so measured a summary of the broad energy patterns of the biosphere and human societies. The summary chapter opens with the solar energy stream, passes on to graph among many cogent graphs the power of prime movers over history, and closes with lively hope for a "grand compromise"—by no means yet inevitable—between the provision of energy enough for a decent quality of human life and the sustenance of the essential functions of the biosphere.

"Everything...can be seen in energy terms," for energy is intrinsic to change, and both life and history are streams of change. No one who knows the critical approach of this Winnipeg author will be surprised by his care to avoid the excesses of the single-variable enthusiasts. The world will not travel along one narrow energy path, whether it be soft or hard. Smil is at pains to explain that maximizing energy use is no national panacea: it does not guarantee prosperity, invulnerability, quality of life, efficiency or even diversity; what it does assure is "higher environmental burdens." China's heavy energy flows—based on a billion tons of coal a year, half of it very hard won indeed—have done less for national advancement than a balanced industrial growth "combined with promotion of the service sector." A careful look at energy flow is distinct from and as deep as any economist's account of GNP in terms of money and much more general. Both are essential tools for thought, not substitutes for it. The final word of the text is the name that Linnaeus claimed for our species: *sapiens.*

The most powerful of about a hundred graphs and tables—a few old drawings are telling, like that of a Victorian noonday traffic jam of people amid horses and carts on London Bridge—sums up our 20th century. It is the record of world harvests and their rising energy subsidies over the years. In 1900 there was almost no direct outside energy investment in agriculture; sunlight almost alone fueled the farm workers and their animals, by courtesy of the green plants.

Half of the world's population, most of the increase that marked this century, is supported by the yield of that subsidy. More than half of that energy goes not to run tractors but to manufacture synthetic fertilizer. The Chinese now spend 2.5 times the American energy subsidy on an average area of crop. "Doing without...would necessitate... a drastic reduction of the global population," even if the rich lands went back to the more direct diet of 1900—wheat flour, potatoes, sugar and a little lard. (We ought not to forget that the world population growth rate peaked about 1970 and has slowly fallen since, a major portent of hope.)

In support of the summary pages, a whole chapter enlarges on the energetics of food; there has been an 80-fold increase this century in the energy provided to augment the sun on every hectare. Even so, less energy is spent in the world's fields than in processing our foods, in packaging, transport and cooking, in flour mills and freezers, and in the little kitchen fires of a billion subsistence farming households. That trend must and will change as our human head count reaches a plateau; there is no source and no sink for much energy multiplication at that scale.

A dozen carefully sequenced chapters spell out the "universal linkage" of energy, in effect quantitative support for each topic of the summary. They open with sunlight and seafloor spreading, go on to trees and soil animals, accumulate various and relevant energy estimates over all of human history. At the last there appear those slow planetary cycles in which currents of carbon, nitrogen and sulfur draw attention for the first time away from energy alone to atomic specificity. Worldwide signs of human interference are clear, after 300 centuries of a human geochemical significance never more than local.

Since Prometheus, our fires have burned the phytomass from green field and forest; our mills have used falling water, wind and fire. By 1600 the great shift to fossil fuel had begun. By 1650

Britain's annual coal output had passed two million tons. The Industrial Revolution came twice: the first time, coal simply meant plentiful fire, the process heat to make salt, soap, glass, beer, even iron. Two centuries later, coal came to mean pistons and steam, the mechanical power to hammer iron on the forge, lift pump rods, turn factory shafts and drive locomotive wheels. Our century is the era of internal combustion. That shift of fuel was remarkably orderly; coal use exceeded that of biomass about 1890, to be overtaken in turn about 1960 by fluid hydrocarbons. Such steady trends encourage theorists but give no real assurance of continuation. Even though for 150 years a simple algebra of growth and substitution has fit the fuel changes beautifully, the world energy system does not in fact have "a schedule, a will, and a clock." We know how much novelty may come with the six o'clock news.

Energy is the physicist's unity within change, but the author, always candid, concedes his "incorrigible fascination with unruly and fuzzy realities in preference to abstract models and dubious generalizations." Consider a few of these. Poor cities under the hot sun, like dense Calcutta, metabolize energy area for area at the same rate as affluent, sprawling and motorized Los Angeles, some 10 or 15 watts per square meter. The meager kitchen stoves and the scarcity of auto engines in the East are offset by the sheer density of Indian dwellings and the small factories hard at work even in residential areas.

Concentrated power has always been critical for particular purposes. It has been reckoned that a work team of 2,400 men was needed to haul into place the heaviest of Inca stone blocks. Those men would produce a sustained power of about 100 watts each, so 250 kilowatts all together, like a big diesel truck. The mean muscle power used during the generations of travail to erect a royal Egyptian pyramid, in the work of masons and haulers, was about enough to build a hundred medieval cathedrals. The largest waterwheel ever built, more than 20 meters in diameter, drained a mine in the Isle of Man as late as 1926 (it is now restored). That wheel yielded 200 kilowatts. Sources of useful energy in the range of gigawatts belong only to this century, apart from the momentary report of cannon.

The pleasure and stimulation of the book come from its critical display of such unruly realities; its importance, from the fact that serious argument on any of these high matters must take into account this army of decisive magnitudes Professor Smil has marshaled.

ESSAY: THE STATE OF SOVIET SCIENCE by Sergei Kapitza

The past 20 years of Soviet science have in no way been the best. Our political scientists describe them as years of stagnation, and this epithet may be applied to science as well. If *perestroika* had occurred in 1968, we would certainly be much better off now.

The stagnation can be illustrated by examples from the field in which I work, particle accelerators. Fifteen years ago we started to build a powerful 800-MeV proton accelerator. This project has still not been completed. Ten years ago we began a 2.5-GeV synchrotron radiation source; we still have not seen light. At present we are building a 3-TeV proton accelerator and collider in a 27-km tunnel. I am not sure we can deliver before the ambitious Superconducting Supercollider accelerator is in place in an 87-km tunnel in Texas.

We have also witnessed the partial failure of our recent Phobos missions to Mars. And of course I have to mention the tragic nuclear accident at Chernobyl. Unfortunately, it was more than an accident. One tends to think it was destined to happen because our society is unable to face the responsibilities of handling advanced technology. This deficiency has only grown worse during the past 20 years.

Of special relevance has been the lack of progress in computing technology, particularly in the development of hardware. Centralized production and tight government control have ceased to exert enough power to control the entire field, as they did in the very successful nuclear and aerospace industries. In fact, we have not in any major way entered the information revolution.

To understand how science is managed in the Soviet Union, one has to look at the four main bodies responsible for research. It is important to keep in mind that funding for all research, which of course comes from the state, has maintained the general proportions among basic research, applied science and the industrial effort of 1:10:100. First, there is the State Committee for Science and Technology. A rather top-heavy governmental body whose chairman is a deputy prime minister, it conducts much of the nonmilitary R&D. Second are the research establishments of the various ministries. These ministries, which are now being disbanded, are best considered as large monopolies, with all the positive and negative points that such organizations develop.

The negatives have definitely contributed to the general crisis of our economy—primarily the lack of competitiveness at home and abroad. A close second, however, is the systematic lack of innovation. Metallurgy, for example, spends only 0.5 percent of its turnover on research. Consequently, half of our steel is still made in open hearth furnaces, and only 15 percent is rolled in continuous casting mills, even though the process was initially developed in our country.

Third is the research at educational institutions. This work is often not of great originality and quality, but because universities do train a large number of scientists and engineers, the country can boast of a rather high level of tertiary education.

Last, but far from least, is the Academy of Sciences of the Soviet Union. Founded in 1724 by one of last decrees of Peter the Great, it is both a learned society and a remarkable network of major institutes, mainly engaged in basic research. Uniting scientists from all over the Soviet Union, the academy is a significant cohesive factor, important at a time of powerful secessionist trends. The academy has produced first-class results and brought forth scientists of very high standards, particularly in mathematics, theoretical physics and chemistry. (Basic research in biology was dealt a deadly blow by Lysenko.)

Divisive and often monopolistic attitudes of the various arms of the scientific establishment have led in most cases to a serious, almost fatal, lack of mobility, both in the connection between research and teaching and in that between research and industrial innovation. Only in military technology have we made the connection effectively. In that field a long line of successes with nuclear weaponry and missiles culminated in *Sputnik*, manned space flights and space exploration. Those achievements have for many years dominated the image of Soviet science. One has to understand that they came about because of a sound basic scientific tradition and a powerful government, which then had full command of immense human and material resources. Today, after losing the cold war, we have to demilitarize our approach and change the way science is supported by the state. That process is difficult and painful because, paradoxically, when science did contribute to the military effort, it could get support for fundamental research.

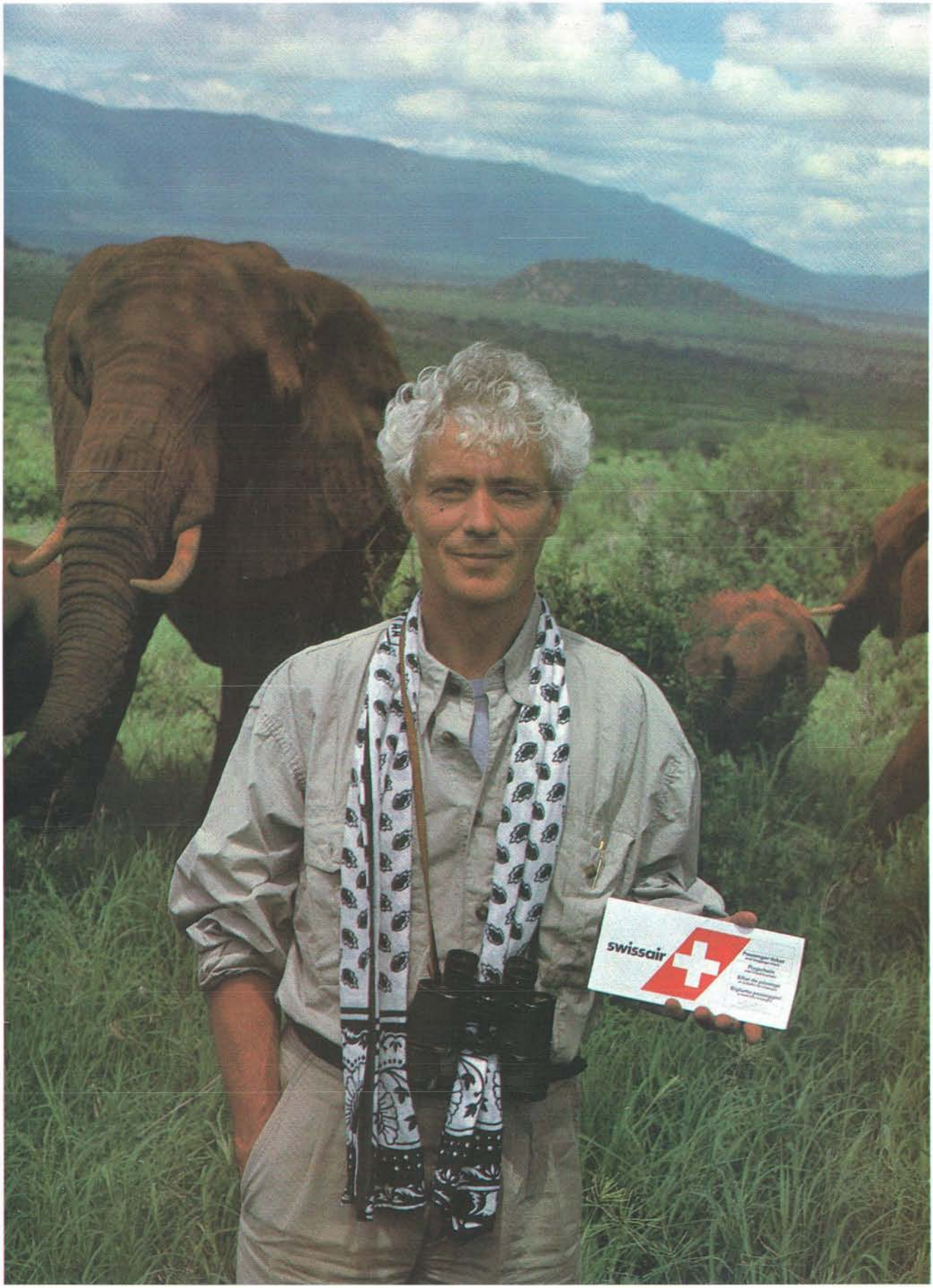
Despite all the failures of the past 20 years, much has changed in recent months. Our country and our science have become open to the world and to ourselves. That is why we are now so self-searching and apprehensive about our state of affairs. Nor should we forget some major assets that we have in the world of science—perhaps the main one being the educational system. With all its drawbacks, its conspicuous lack of funding and a certain decline in standards—problems that, I may say, are not unique—we still produce dedicated and relatively well trained manpower. The point is to use it efficiently. Here much depends on the new degrees of freedom in the economy.

Another asset is our tradition of integrative studies of nature. Today, with urgent environmental problems looming on the world's common horizon, this capacity for interdisciplinary research is gaining recognition. In basic research, if we overcome the tendency to commercialize science—and such overzealous marketeers also exist in our part of the world—we will manage, with luck even expand, our research efforts. Recently a presidential decree has given support to the Academy of Sciences, providing funds and protection for research. Unfortunately, the difficulties with funding, especially those caused by the nonconvertibility of our currency, may severely hamper international collaboration. Here help from abroad would be a sensible and effective way of supporting the momentous changes taking place in our country.

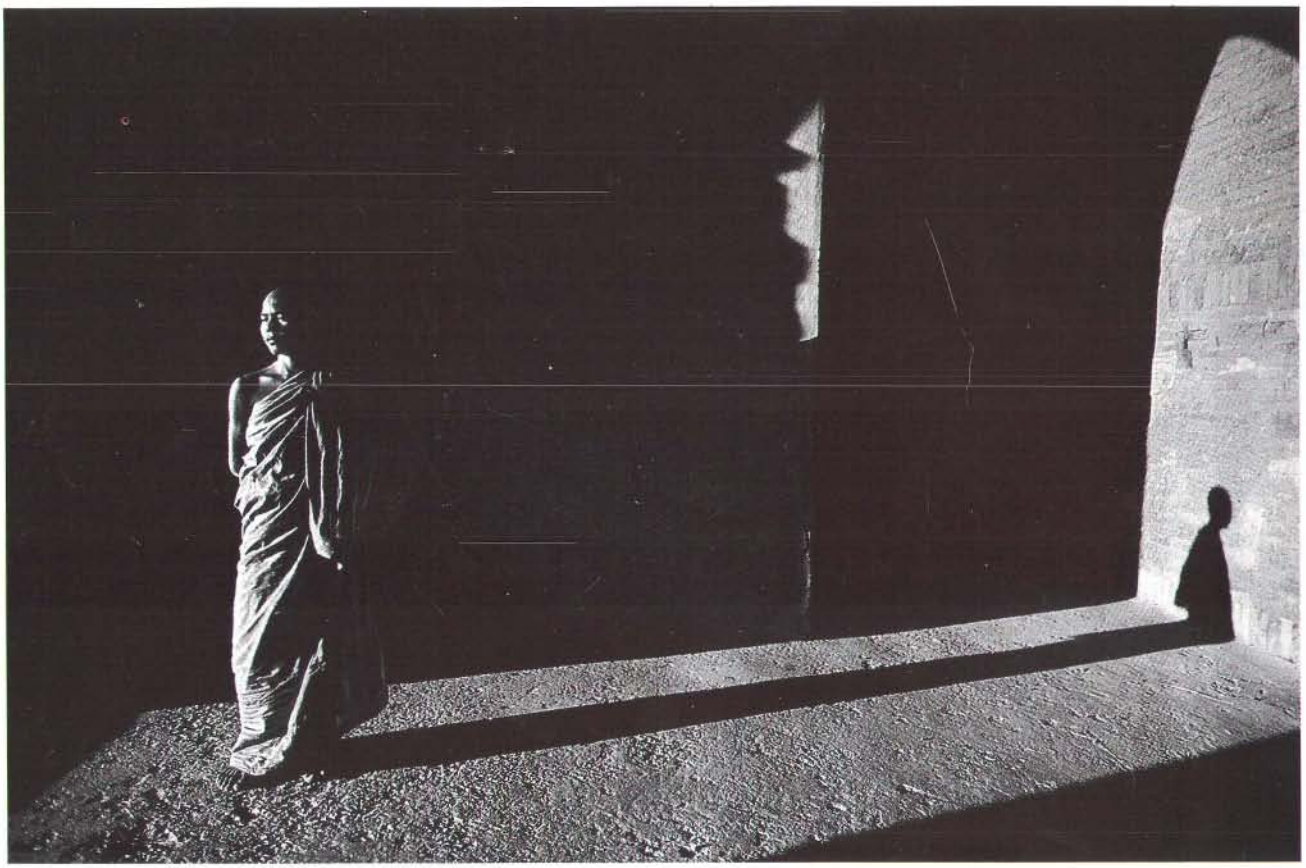
Scientists en masse belong to the constituency of Gorbachev. The success of his policies will determine the future of science in the Soviet Union. In the long run, science may be our most valuable asset, for one can run out of oil but not out of brains, which tend to multiply as long as they are given a chance.

SERGEI KAPITZA, a physicist and a member of the Soviet Academy of Sciences, is editor of the Russian edition of SCIENTIFIC AMERICAN.

Swissair Customer Portrait 88: W. Bobby Gschwend, director of a Safari enterprise, Mombasa, photographed by Alberto Venzago.



CONTEMPLATION



LEICA CAMERA GMBH, OSKAR-BARNACK-STRASSE 11, D-6336 SOLMS

MADE BY LEICA

The thrill of photography often lies in seeing without being seen. That means acting quickly, discreetly, and sensitively. For this, the LEICA M6 is just right. The compactness, the barely audible shutter release, the exemplary fine mechanics, and a lens range second to none, crowned by the 50 mm f/1 Noctilux, lend this camera its excellent quality and unique character. And give the photographer access to a new, fascinating dimension of photography. Leica photography.

Leica

The freedom to see.